MIT/LCS/TR-533

MODELLING SPEAKER VARIABILITY AND IMPOSING SPEAKER CONSTRAINTS IN PHONETIC CLASSIFICATION

Partha Niyogi

February 1992

92-28932

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	Feb. 1992	

4. TITLE AND SUBTITLE

Modelling Speaker Variability and Imposing Speaker Constraints in Phonetic Classification

5. FUNDING NUMBERS

6. AUTHOR(S)

Partha Niyogi

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Massachusetts Institute of Technology
Laboratory for Computer Science
545 Technology Square
Cambridge, MA 02139

8. PERFORMING ORGANIZATION REPORT NUMBER

MIT/LCS/TR 533

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

DARPA

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

N00014-89-J-1332

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT

12b. DISTRIBUTION CODE

13. ABSTRACT (Maximum 200 words)

This thesis deals with intra-speaker correlation analyses of speech sounds, and the possible utilization of this correlation to speech recognition. Current approaches to phonetic classification, regardless of whether they use context-dependent or -independent models, achieve classification based on locally optimum criteria. They make no fundamental assumption about the fact that the same vocal tract is used to make all the phonemes in an utterance. Thus, for example, a system may classify one sound in the beginning of an utterance as an /s/ belonging to a long vocal tract, while inappropriately classifying another sound in the same utterance as an /ʃ/ belonging to a short vocal tract. Clearly the different phonemes of an utterance are correlated. Hence there is a set of speaker-specific constraints that can be imposed among all sounds in an utterance, and phonetic decoding should be accomplished by exploiting these constraints.

(con't.)

14. SUBJECT TERMS	15. NUMBER OF PAGES
	108
	16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT

13.A   To investigate this approach, we formulated the problem mathematically into four paradigms, each incorporating a different amount of speaker-specific constraints. We obtained empirical results on a constrained task of speaker-independent vowel classification. Controlled studies of the performance of the different paradigms were conducted. Parameters such as number of training and test tokens, classifier used, methods of clustering speakers into representative speaker groups were varied systematically. An attempt was made to understand the conditions under which imposition of speaker constraints led to potential improvement in recognition accuracy. Later, we expanded our task to classification of all phonemes in American English and found that improvements in performance due to speaker constraints were maintained.

Modelling Speaker Variability and Imposing Speaker Constraints in Phonetic Classification

by

Partha Niyogi
B.Tech. Indian Institute of Technology, New Delhi.
(1989)

Submitted to the Department of
Electrical Engineering and Computer Science
in partial fulfillment of the requirements
for the degree of

Master of Science
at the

Massachusetts Institute of Technology
December, 1991

Signature of Author..
Department of Electrical Engineering
May, 1990

Certified by...
Victor W. Zue
Principal Research Scientist,
Department of Electrical Engineering and Computer Science
Thesis Supervisor

Accepted by..
Arthur C. Smith
Chair, Department Committee on Graduate Students

Modelling Speaker Variability and Imposing
Speaker Constraints in Phonetic Classification

by

Partha Niyogi

Abstract

This thesis deals with intra-speaker correlation analyses of speech sounds,
and the possible utilization of this correlation to speech recognition. Current
approaches to phonetic classification, regardless of whether they use context-
dependent or -independent models, achieve classification based on locally
optimum criteria. They make no fundamental assumption about the fact
that the same vocal tract is used to make all the phonemes in an utterance.
Thus, for example, a system may classify one sound in the beginning of
an utterance as an /s/ belonging to a long vocal tract, while inappropriately
classifying another sound in the same utterance as an /ʃ/ belonging to a short
vocal tract. Clearly the different phonemes of an utterance are correlated.
Hence there is a set of speaker-specific constraints that can be imposed among
all sounds in an utterance, and phonetic decoding should be accomplished
by exploiting these constraints.

To investigate this approach, we formulated the problem mathematically
into four paradigms, each incorporating a different amount of speaker-specific
constraints. We obtained empirical results on a constrained task of speaker-
independent vowel classification. Controlled studies of the performance of the
different paradigms were conducted. Parameters such as number of training
and test tokens, classifier used, methods of clustering speakers into represen-
tative speaker groups were varied systematically. An attempt was made to
understand the conditions under which imposition of speaker constraints led
to potential improvement in recognition accuracy. Later, we expanded our
task to classification of all phonemes in American English and found that
improvements in performance due to speaker constraints were maintained.

Acknowledgements:

To Victor Zue, my thesis advisor, for creating an outstanding research environment, for sharing with me his quest for perfection, for professional and personal support, and for treating me as a friend;

To the members of the Spoken Language Systems Group, for assistance, and teaching me about speech;

To my officemates, Jeff and Lee, for help with UNIX/C/Splus/Statistics, for endless discussions, and for friendship;

To all those people, here and in India, for providing me with a life;

And to my family, for constant support and encouragement, for love and affection, and countless other things;

I offer my deepest gratitude.

Contents

List of Figures

8

9

List of Tables

Chapter 1

Introduction

1.1 Background

Over the last decade, there has been an increased interest in developing speech recognition systems. The goal of such a system is to take as its input the acoustic waveform uttered by a human and produce the corresponding string of words. It tries to achieve an optimal mapping between the acoustic signal and a lexical representation.

This mapping from the acoustic to the lexical domains is one to many, and very often a unique, exact solution does not exist. Various assumptions need to be made about the nature of the signal and the underlying physical processes of speech production and perception. One such assumption is that different sounds produced by a speaker are uncorrelated and so the mapping from sound to lexical units can be done independently for different sounds. This thesis argues that such an independence assumption is not valid, and further develops algorithms to perform the mapping of the different sounds to the lexical units, jointly, rather than individually.

Speech recognition is very difficult because of the enormous variability in the speech signal. This variability may be due to many reasons. For exam-

ple, the acoustic realization of a certain phoneme depends on its context, i.e., the phonemes which lie near it[1]. As an example of these context dependencies, the realization of a vowel next to an /r/ would have different acoustic characteristics[2] from that of the same vowel near a nasal, and both would be different from a canonic version of the vowel. In addition to context, speaker characteristics also account for some of the variability. Speaking rate, style, stress patterns all affect the speech signal. Furthermore, there are fundamental factors like the size and shape of the vocal tract which play an important role. Shown in the spectrograms of Figure 1.1, are two examples of the same phonetic string, one uttered by a male and one by a female. Notice how the female has higher formants in all her vowels compared to the male. This is because the female had a shorter vocal tract and the length of the tract is inversely related to the values of the formants (as a first order approximation).

1.2 Some Issues of Importance in Speech Recognition

As has been described in the previous section, speech recognition is a very difficult problem. Consequently, scientists have tackled it at various levels of complexity, and many kinds of speech recognition systems have been developed. These systems differ from each other in the nature of the recognition task, and the algorithms used to perform it. For example, some systems try to recognize isolated words only, others try to recognize connected speech.

---

[1] Phonemes are the basic linguistic units which make up a language. A phoneme is the basic contrastive sound unit and several phonemes concatenated together constitute a word.

[2] One measure of acoustic characteristics could be formant values. Formants are resonant frequencies of the vocal tract.

13
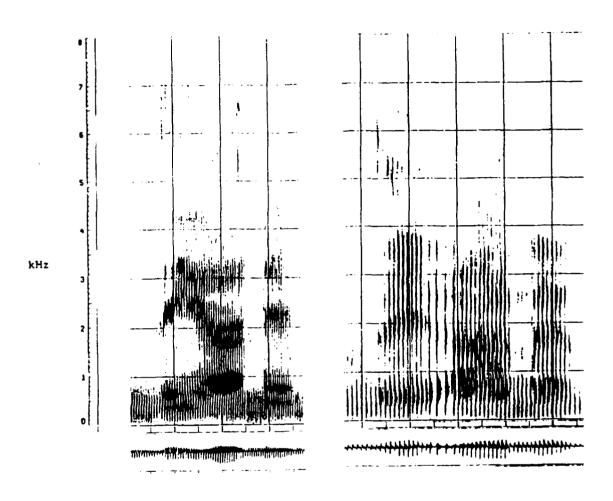
Figure 1.1: Spectrogram of same phonetic string /mȫιʌnɛ/ uttered by the female on the left and male on the right

Recognizers may differ depending upon whether they handle large or small vocabulary sizes, multiple speakers or single speakers, etc. Even if two recognition systems work on similar problems, they might use different recognition methodologies. For example, researchers have tackled the problem of continuous speech recognition in several different ways. Some people [28] attempt to segment the speech signal into acoustically homogeneous segments, assign each segment an ordered list of likely phonetic labels, and then choose a phonetic transcription for the entire acoustic signal subject to an optimality criterion. Another very common technique is to model the acoustic utterance as the output of a Markov process with models for individual phonemes connected together [16] according to language constraints[3]. Here no segmentation of the signal is required and the sentence is recognized on the basis of which combination of models best fit the acoustic waveform.

Whatever the problem one chooses to work on, and whatever the recognition framework one uses, there are two issues which are relevant across all multi-speaker systems at most levels of complexity. Firstly, it is necessary to model the speech signal closely and account for its variabilities. Secondly, for superior performance, it is preferable that the system adapt in some way to test speakers. These issues are particularly noteworthy because they are related in part to the ideas of this thesis.

1.2.1 Modelling the Variability in Speech

There are various sources of variability in speech. Some of the variability is due to inter-speaker differences. Rabiner [22] developed an isolated-word, speaker-independent speech recognition system by clustering speakers, and forming multiple reference templates for each word against which the test

---

[3]This technique known as Hidden Markov Modelling (HMM) is very popular today.

word was compared. A nearest neighbor decision scheme was used. The clustering of speakers into groups helped to take care of speaker variability to some extent. More recently, Murveit et al. [20] have used parallel male and female models in an HMM based continuous speech recognition system used in a speaker-independent manner. This enabled them to decrease the word error-rate from roughly 5.2% to 4.5% on DARPA's February 1989 speaker-independent test set for the Resource Management task using the standard perplexity 60 word-pair grammar.

Another source of variability in the acoustic realization of phonemes is its phonetic environment. Triphone modelling, first introduced by researchers at IBM and BBN [23], account for contextual variation of phonemes by using different models depending on the left and right context. K.F. Lee in his SPHINX continuous speech recognition system [16] [17] made use of generalized triphones which were obtained by collapsing some contexts.

Often the training speech data are assumed to be distributed in a Gaussian fashion. This is usually a faulty assumption. Of late, C.H. Lee and others [15] at AT&T Bell have tried to use a mixture of densities, usually Gaussian, to characterize the data which was represented earlier by a single Gaussian. This allows for closer approximation of the training data and improves performance.

1.2.2   Speaker Adaptation

A lot of effort has been spent on developing algorithms for speaker adaptation. This usually involves collecting a small amount of training speech from the test speaker and then appropriately updating the models based on his or her speaker characteristics. These updated models are then used to recognize more speech in the testing phase. Lasry and Stern [26] developed a methodology for updating the mean and covariance for the acoustic representation for

some sounds on the basis of the training samples of not only those sounds, but of other sounds also produced by the same speaker in the adaptation phase. Various techniques have been used to map the templates (models) for a reference speaker to that of the input speaker. Choukri and Chollet [3] used Canonical Correlation Analysis to perform a spectral transformation from reference to test speaker. A probabilistic spectral transformation has been suggested by [5]. Shikano [25] developed algorithms using vector quantization codebook mapping.

1.2.3 Discussion

Some of the above schemes take labelled speech in the adaptation phase and compare it with the same utterances from the reference speaker in performing spectral transformations. Models for a particular sound are thus updated on the basis of examples of only that sound uttered by the test speaker. This does not explicitly exploit correlations between the different sounds produced by the same speaker. Furthermore, once the adaptation phase is over, there is usually no further attempt to update the models in the test phase. As a matter of fact, when recognizing unlabelled speech, many of the above-mentioned techniques are locally optimal in that they map the acoustical to lexical domains segment by segment. For a phonetic classification task, this means that even if the test speaker has uttered a lot of phonemes, each phoneme is classified independently. For word recognition, each word uttered by the test speaker is recognized independently and in continuous speech recognition, different parts of a sentence are assumed independent and treated as so.

While such independence assumptions allow for computationally tractable solutions, they again do not explicitly exploit correlations between different sounds produced by the same speaker. Lasry and Stern make use of these correlations only in the adaptation phase. In the testing phase, all the different

17

tokens are treated independently.

The same can also be said of those schemes which don't operate in a speaker-adaptive mode but have speaker models instead. Rabiner [22] classifies test words one at a time and hence no speaker constraints are imposed on test speech.

1.3 Thesis Overview

The goal of this thesis is to try and explore various ways in which these correlations between different sounds could be exploited for phonetic recognition. We examine several ways to model the speaker variability, and then in the recognition phase, we try to enforce the constraint that different tokens produced by the same speaker are correlated and that the acoustic-to-lexical mapping should be performed jointly or in a globally optimal way.

Chapter 2 of this thesis provides some evidence that different sounds produced by the same speaker are correlated. An approach based on linear regression has been used to characterize some of these correlations between vowel pairs. Correlation of sounds with gender of the speaker is also demonstrated. This is followed by a mathematical formulation of different paradigms of classification which enforce the speaker constraint in different ways and to different degrees. A few toy examples illustrate feasibility of the ideas.

Chapter 3 compares and contrasts the different models with the baseline under different conditions for a specific task of vowel classification of eight vowels. This is an implementation of the generalized theory developed. Various issues involving the engineering trade-offs between improvement in classification accuracy, model assumptions and computational complexity are investigated and resolved.

Chapter 4 discusses the implementation of the best model under best operating conditions on a larger set of phonemes in order to see if the results generalize.

Chapter 5 provides the summary and concludes the thesis by reiterating most of the important results.

Chapter 2

Mathematical Formulation

2.1 Evidence for Correlation of Speech Sounds Produced by the Same Speaker

We have mentioned earlier that the speech signal has a vast amount of variability. A lot of this variability is due to inter-speaker differences. Speaking rate, stress patterns, pitch, size and shape of the vocal tract are amongst the many factors which affect a speaker's acoustic signal. However, these speaker characteristics are likely to remain consistent over all sounds uttered by that speaker. After all, the different sounds produced by him or her have been produced by the same sound-producing apparatus and they should hence be correlated to some degree.

In this thesis we intend to exploit these correlations and develop recognition algorithms which do not classify different sounds produced by the same speaker individually but rather do so jointly. This effectively enforces some acoustic constraints particular to that speaker. Before we proceed to develop the mathematical framework for such a task, we intend to provide some evidence that different sounds are indeed correlated.

As an example, let us look again at the two spectrograms in Figure 1.1. One is a male speaker and the other is a female speaker. Notice in particular the formant values for each speaker. The female has a shorter vocal tract and according to the acoustic theory of speech production, has higher formant values. This is so for all vowels produced by the female. So, for example, comparing just the /ʌ/ for each of the two speakers gives us a rough idea of how their /ɪ/'s would compare. Similarly the fundamental frequency of the female speaker is higher throughout the utterance. Knowledge of the acoustic character of some parts of the utterance helps us to predict the acoustic character of other parts.

To quantify this correlation over a larger number of speakers, we conducted an experiment using the TIMIT corpus [14]. This corpus was designed jointly by researchers at MIT, TI and SRI. It consists of a total of 6,300 sentences from 630 speakers, representing over 5 hours of speech material, and was recorded by researchers at TI. Each speaker in the TIMIT corpus recorded 10 sentences drawn from three different sources as follows. Each speaker read two sentences (common for all speakers), designated as SA sentences which were designed at SRI in order to compare dialectical and phonological variations across speakers. Five sentences, designated as SX were drawn from a set of 450 sentences designed at MIT. The remaining three sentences for each speaker, designated as SI sentences, were selected from the Brown corpus [13] at TI. Each SI sentence was unique and differed across speakers.

In our experiment we selected 396 speakers from this corpus and chose one SA sentence per speaker. This was the same for all the speakers and had the following orthographic transcription - "She had your dark suit in greasy wash water all year". We selected the /æ/ from the word "had" for each speaker with /h/ and /d/ as its left and right context. Similarly we selected

21

the /iʸ/ with /ʃ/ and /h/ as its left and right context for each speaker. Thus we had 396 pairs of /æ/ and /iʸ/ for each speaker. The measurement made on the speech signal was the first autocorrelation coefficient, $r[1]$, defined as:

$$r[1] = \sum_n s[n]s[n+1] \tag{2.1}$$

where $s[n]$ is the speech signal. It can easily be shown that

$$r[1] = \frac{1}{2\pi} \int_{-\pi}^{\pi} \|S(e^{jw})\|^2 \cos(w)dw \tag{2.2}$$

Thus $r[1]$ measures a weighted spectral average. The spectrum is weighted by a cosine function. It weights the low-frequency energies positively and the high-frequency energies negatively. In actuality, the short time autocorrelation coefficient was calculated on a frame-by-frame basis using a sliding Hamming window of length 400 samples which was moved 80 samples at a time. The sampling rate is 16 KHz, so each frame represents 5 ms of speech. The value of $r[1]$, averaged over the frames which made up the middle-third of each vowel token was used as the measurement on each vowel. /iʸ/'s are more front[1] than /æ/'s and consequently have higher second and third formants. Correspondingly they usually have lower values for this measurement. We would expect that those speakers who had low $r[1]$ values for their /iʸ/'s presumably had higher formants in general and consequently would also have low $r[1]$ values for their /æ/'s. Shown in Figure 2.1 is a plot of the 396 /iʸ/-/æ/ pairs. A certain degree of correlation is observed in that there is an increase in the $r[1]$ value for the /æ/'s with an increase in that of the /iʸ/'s, but the data is very noisy. To make this trend more visually dramatic, we removed some of the variability by averaging. We divided the

---

[1]This means that the tongue body is fronted and the pharyngeal cavity is wider and less obstructed while uttering the /iʸ/
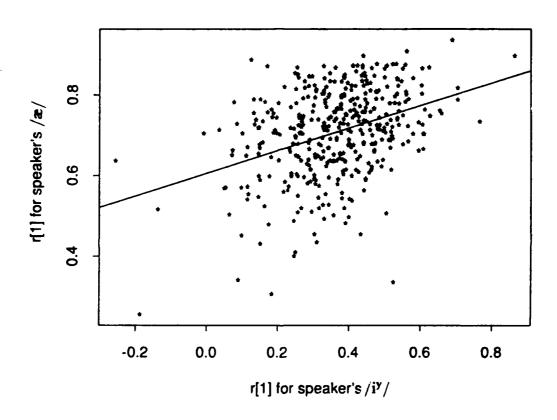
Figure 2.1: Plot of r[1] values for /i^y/'s and /æ/'s of each speaker. Each point represents a speaker. The x-coordinate of the point is the r[1] for his/her /i^y/ and the y-coordinate is the r[1] for his/her /æ/.

Mean r[1] for speaker group's /æ/

0.76 0.72 0.68 0.64

0.1   0.2   0.3   0.4   0.5   0.6

Mean r[1] for speaker group's /iʸ/
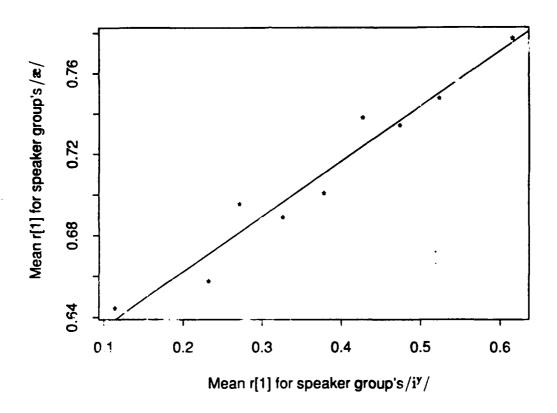
Figure 2.2: Plot of average $r[1]$ values for /iʸ/'s and /æ/'s of each speaker group. Each point represents a speaker group. The x-coordinate of the point is the mean $r[1]$ for /iʸ/'s of all speakers in the group. The y-coordinate is the mean $r[1]$ for the group's /æ/'s.

speakers into 9 equal groups according to the value of $r[1]$ for their /iʸ/'s. In other words, we formed an ordered list of the speakers by arranging them in increasing value of $r[1]$ for their /iʸ/'s. The first 44 speakers then formed group 1, the second 44 formed groups 2, and so on. We thus grouped the speakers into 9 groups purely on the basis of $r[1]$ values for their /iʸ/'s with members of each group having similar acoustic measurements for this value. We now plot in Figure 2.2, the values of the average $r[1]$ for /iʸ/ and /æ/ for each speaker group and the trend is clearer. Those groups who have high

24

average values for $r[1]$ for their /iʸ/'s also have high average measurements for $r[1]$ for their /æ/'s. Clearly there is a correlation. A line of least-squares fit is plotted.

Although, Figures 2.1 and 2.2 suggest that the /æ/'s and /iʸ/'s for the 396 speakers are correlated, we would like to quantify and test for this correlation. Linear regression [21] allows us to do it. Our data consists of 396 (x, y) pairs where x is the value for $r[1]$ for that speaker's /iʸ/ and y is the value of $r[1]$ for that speaker's /æ/. We try to fit a linear model of the form

$$Y_i = \alpha_1 + \beta x_i + \epsilon_i \tag{2.3}$$

where ϵ_i's are all normally distributed, $N(0, \sigma^2)$, and are independent. Clearly if $\beta = 0$, then there is no relationship between a speaker's /æ/ and /iʸ/. We predict y using our linear model and define the sum of squares of the errors over all $n = 396$ speakers to be

$$H(\alpha_1, \beta) = \sum_{i=1}^{n} [y_i - \alpha_1 - \beta x_i]^2 \tag{2.4}$$

We choose α_1, β to minimize $H(\alpha_1, \beta)$. The optimal values can be denoted as $\hat{\alpha}_1, \hat{\beta}$. We can actually test for the hypothesis $H_0 : \beta = 0$ against $H_1 : \beta \neq 0$. To do this we need to calculate a T-statistic [10] according to

$$T_1 = \frac{\hat{\beta}}{[n\widehat{\sigma^2}/[(n-2)\sum_1^n (x_i - \bar{x})^2]]^{1/2}} \tag{2.5}$$

This T-statistic has $n - 2$ degrees of freedom. For our case of 396 speakers, we obtain $\beta = 0.27$ and $T_1 = 7.93$ which is significant at the 0.005 level. This indicates that β is non-zero. In other words, knowledge about a speaker's /iʸ/ helps us to predict his or her /æ/. (Of course, in this case by simply reversing the (x, y) tuples, we can do equally well in predicting the /iʸ/ from

the /æ/.) A measure of fit for this model is the Coefficient of Determination (R) [21] defined by

$$R^2 = \frac{\sum (\hat{y}_i - \bar{y})^2}{\sum (y_i - \bar{y})^2} \tag{2.6}$$

where $\hat{y}_i$ is the predicted value of y_i for each x_i according to our model. This measure indicates the proportion of variability in the y's explained by the model. We obtained a value of 0.137 for R which is very low. This is hardly surprising since our measurements were extremely simple, we had only one token per speaker (rather than an average of many which would have added more robustness) and our model was a simple linear one. The purpose of this experiment is not to try and account for all the variability in a phoneme by knowledge of another but to show that we can account for some of it by simple correlation. This simple experiment indicates that the /i^y/'s and /æ/'s for the speaker are correlated. Obviously more complicated models and more complicated measurements would help us capture these correlations better. Also from Figure 2.1, we get an idea of the variability amongst the speakers. With this as motivation, we will now develop the mathematical framework for our task.

2.2   Development of the Mathematical Framework

2.2.1   Conceptual Formulation

In the earlier section we have seen some evidence of inter-speaker acoustic differences and intra-speaker acoustic correlations of different sounds. Closer modelling of these factors might lead to potential improvement in classification performance. Figure 2.3 indicates a simple one-dimensional two-class

26
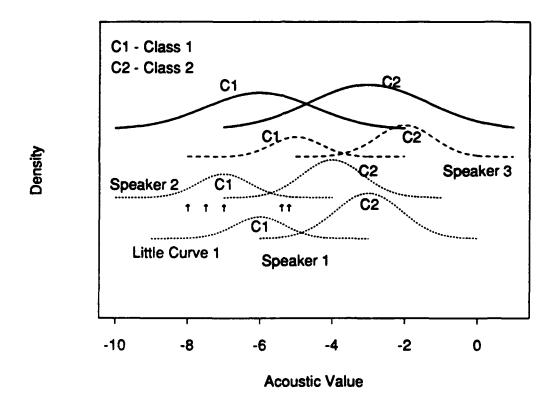
Figure 2.3: A simple illustration of the multiple-speaker scenario for two sound classes C1 and C2. The solid lines are overall distributions by pooling all speakers.

situation to illustrate our viewpoint. The x-axis is the acoustic value and the y-axis is proportional to probability density. Thus the distribution labelled Little Curve 1 is that of the acoustic value given Class 1 and Speaker 1. In the figure shown, there are only 3 speakers or speaker types. The solid lines indicate the overall distribution for each class by pooling all the speakers together into one group. This figure represents our general model of speaker variability. The different speakers lie in different regions in acoustic space. Moreover, the different sounds produced by them (in this case each

speaker produces only two kinds of sounds) are correlated. In the example, if a speaker has a higher mean for sounds corresponding to Class 1, he or she would have a higher mean for sounds corresponding to Class 2. Suppose Speaker 2 has produced 5 tokens, as indicated by the 5 arrows whose x-coordinate indicates the value of the acoustic vector. Classifying these tokens using the broad pooled-speaker distributions would be suboptimal. As is clear from the figure, the broad distributions have greater variance, poorer resolution and hence result in a higher error-rate. In our specific example, we would probably have classified all 5 tokens as belonging to Class 1. However, looking at the acoustic distributions of Speaker 2, we intuitively feel that this is not so. At the same time, using the speaker-specific distributions for a different speaker is suboptimal too. This is seen by applying the distributions of Speaker 3 to the classification task in which case all our 5 tokens would again be classified as belonging to Class 1. Classification using the speaker-specific distributions of Speaker 2 is optimal. If the right speaker-specific curves can't be used, we would at least like to impose the constraint that all these tokens are produced by the same speaker and correspond to a distribution pair. Thus if we classify the first three tokens from the left as belonging to Class 1, it should provide an estimate of the acoustic nature of Class 1 tokens produced by the speaker. Making use of our premise that sounds belonging to different classes are correlated if produced by the same speaker, we would presumably have developed estimates of Class 2 tokens for the same speaker. Consequently, the two tokens on the right would then be classified as belonging to Class 2. In other words, there are two things we would like to do

- Decompose the overall population of speakers into speaker-specific models to capture inter-speaker variability.
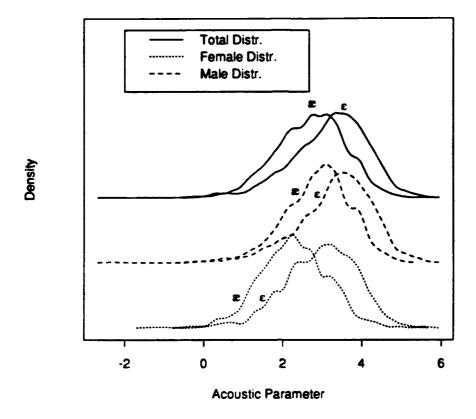
Figure 2.4: The density distribution of a typical acoustic parameter for the vowels /æ/ and /ɛ/. The top curves represent pooled data, whereas the middle and bottom curves represent the data for male and female speakers separately.

• Classify tokens produced by the same speaker jointly so that we can exploit intra-speaker correlations of different sounds.

The following section gives some mathematical rigor to these ideas. Figure 2.4 shows distributions computed from real data and demonstrates the closeness of our model to reality. In this case, the two classes are the phonemes /æ/ and /ɛ/ and there are two speaker types - males and females. The curves are obtained by pooling together tokens from male and female speakers respectively from the TIMIT corpus. Each vowel token was represented by a spectral average. The acoustic space was further rotated using discriminant functions, and the acoustic parameter plotted is the first

29

discriminant function of the spectral average. Decomposing the overall distributions into speaker-specific ones thus seems a valid thing to do.

2.2.2 Mathematical Formulation

We start by introducing the following set of notations:

n is the number of linguistic classes (e.g. phonemes, triphones, or words), labelled as $\{w_i; i = 1, .., n\}$,

N is the number of speakers, labelled as $\{S_i; i = 1, .., N\}$,

$\vec{x}$ is the acoustic vector produced by a speaker when uttering a certain class,

$p(\vec{x}|S_i, w_j)$ is the probability density of the acoustic vector given speaker i and class j,

$p(w_j)$ is the *a priori* probability that a speaker utters class w_j. We assume that this is independent of the speaker, i.e., $p(w_j|S_i) = p(w_j)$, and

$p(S_i)$ is the *a priori* probability that any given test speaker is the i-th speaker.

Let us assume that we have in hand a set of acoustic tokens, $\{\vec{x}_i; i = 1, .., L\}$ produced by a given speaker. These tokens could, for example, correspond to different segments of a sentence. Our task is to classify each of the tokens into one of the n linguistic classes. Specifically, we want to determine the optimum classification of $\vec{x}_j$ as C_j, for all j, where $C_j \in \{w_i; i = 1, .., n\}$. C_j is thus a variable which can take on any one of n values and we want to choose the optimal one, according to an optimality criterion. The most straightforward procedure would be to pool the acoustic data for all speakers

30

into a single distribution, $p(\vec{x}|w_i)$, as illustrated in the top curves of Figure 2.4. Traditionally, the unknown tokens are classified independently, i.e., we independently choose the class C_j that maximizes $p(C_j|\vec{x}_j)$. This is classical Bayesian classification, assuming independence between tokens. It is equivalent to choosing the classes $C_1, .., C_L$ using the following criterion:

$$\max_{C_1,...,C_L} \prod_j p(C_j|\vec{x}_j) \tag{2.7}$$

Using Bayes rule,

$$\max_{C_1,...,C_L} \prod_j \frac{p(C_j)p(\vec{x}|C_j)}{p(\vec{x}_j)} \tag{2.8}$$

As $p(\vec{x}_j)$ is independent of C_j, we can ignore it in Eq. (2.8) and instead carry out the following equivalent maximization.

$$\max_{C_1,...,C_L} \prod_j p(C_j)p(\vec{x}_j|C_j). \tag{2.9}$$

In reality, the acoustic models of a population are speaker-dependent, as illustrated by the middle and bottom curves in Figure 2.4. By decomposing the overall models into male and female counterparts, for example, we can get tighter distributions, thus leading to potential performance gain. More generally,

$$p(\vec{x}_j|w_i) = \sum_{k=1}^{N} p(S_k)p(\vec{x}_j|S_k, w_i), \text{ and} \tag{2.10}$$

$$p(\vec{x}_j|S_i) = \sum_{k=1}^{n} p(w_k)p(\vec{x}_j|S_i, w_k). \tag{2.11}$$

These equations suggest that $p(\vec{x}_j|w_i)$ and $p(\vec{x}_j|S_i)$ can be interpreted as mixtures of densities. The basic components of all the mixtures are $p(\vec{x}|w_i, S_j)$ which corresponds to the speaker-specific distributions in the figure. We could, therefore. classify the tokens collectively by imposing speaker-specific

31

constraints. Depending upon the degree to which we impose such constraints, we can obtain four different classification paradigms.

2.2.3 Paradigm 1: Incorporating Speaker-Specific Models

Assuming that there are N speakers, each with a different distribution, by substituting Eq. (2.10) into Eq. (2.9), we obtain:

$$\max_{C_1,...,C_L} \prod_{j=1}^{L} [p(C_j) \sum_{k=1}^{N} p(S_k)p(\vec{x}_j|S_k,C_j)]. \tag{2.12}$$

In the above equation, we have introduced speaker-specific models. However, the assignment of the classes is still achieved one token at a time, independen' of one another. This will serve as a suitable baseline for comparison.

2.2.4 Paradigm 2: Incorporating Speaker-Specific Constraints Without Speaker Classification

An alternative method of incorporating speaker specific constraints can be found by noting that Eq. (2.10) can be rewritten as:

$$\max_{C_1,..C_L} [\sum_{i1=1}^{N} p(C_1)p(S_{i1})p(\vec{x}_1|S_{i1},C_1)]..[\sum_{iL=1}^{N} p(C_L)p(S_{iL})p(\vec{x}_L|S_{iL},C_L)] \tag{2.13}$$

or equivalently,

$$\max_{C_1,...,C_L} [\prod_{j=1}^{L} p(C_j)] \sum_{i1=1}^{N} .. \sum_{iL=1}^{N} [p(S_{i1})p(\vec{x}_1|S_{i1},C_1)..p(S_{iL})p(\vec{x}_L|S_{iL},C_L)] \tag{2.14}$$

Notice that in Eq. (2.14), terms like $p(\vec{x}_1|S_{i1},C_1)$ and $p(\vec{x}_L|S_{iL},C_L)$ are multiplied to give a finite contribution. Since in general $i1 \neq .. \neq iL$, the term

inside the square brackets in Eq. (2.14) measures the likelihood of $\vec{x}_1, .., \vec{x}_L$ being produced by speakers $S_{i1}, .., S_{iL}$. This is an irrelevant contribution and should be eliminated since the tokens could not have been produced by different speakers. These cross terms exist in Paradigm 1 because of the independence assumption. Hence, it is meaningful to remove that assumption and instead maximize the following:

$$\max_{C_1,...,C_L} p(C_1, .., C_L | \vec{x}_1, .., \vec{x}_L) \tag{2.15}$$

This is equivalent to

$$\max_{C_1,...,C_L} \sum_{i=1}^{N} p(S_i, C_1, .., C_L | \vec{x}_1, .., \vec{x}_L) \tag{2.16}$$

$$\max_{C_1,...,C_L} \sum p(S_i | \vec{x}_1, .., \vec{x}_L) p(C_1, .., C_L | \vec{x}_1, .., \vec{x}_L, S_i) \tag{2.17}$$

$$\max_{C_1,...,C_L} \sum p(S_i | \vec{x}_1, .., \vec{x}_L) \frac{p(C_1, .., C_L | S_i) p(\vec{x}_1, .., \vec{x}_L | C_1, .., C_L, S_i)}{p(\vec{x}_1, .., \vec{x}_L | S_i)} \tag{2.18}$$

$$\max_{C_1,...,C_L} \sum \frac{p(S_i)}{p(\vec{x}_1, .., \vec{x}_L)} p(C_1, ..C_L) p(\vec{x}_1, .., \vec{x}_L | C_1, .., C_L, S_i) \tag{2.19}$$

$p(\vec{x}_1, .., \vec{x}_L)$ can be neglected in the maximization process. Furthermore, we assume that for a particular speaker, the probability that he or she utters class w_i is independent of all other classes he or she has uttered in the past or will utter in the future. Moreover context dependencies in acoustics (as we discussed in Chapter 1) have also been ignored. In effect, only within a particular speaker can the tokens be treated as independent. Thus

$$p(C_1, .., C_L) = p(C_1) p(C_2)..p(C_L) \tag{2.20}$$

$$p(\vec{x}_1, .., \vec{x}_L | C_1, ..., C_L, S_i) = \prod_{j=1}^{L} p(\vec{x}_j | C_j, S_i) \tag{2.21}$$

33

$$\max_{C_1,...,C_L} \sum_{i=1}^{N} p(S_i) \prod_{j=1}^{L} p(C_j)p(\vec{x}_j|S_i,C_j) \qquad (2.22)$$

Unlike in Eq. (2.14), all terms in Eq. (2.22) involving products from different speakers have been eliminated.

2.2.5 Paradigm 3: Incorporating Speaker-Specific Constraints With Speaker Classification

An alternative approach to Paradigm 2 is to classify, for each speaker, the tokens according to that speaker's distributions and measure its likelihood. We can then choose, among the results from all speakers, the most likely answer. This is equivalent to choosing a speaker and a classification based on that speaker's distributions which is most likely given the tokens. Mathematically,

$$\max_{S_i,C_1,...,C_L} p(S_i, C_1, .., C_L|\vec{x}_1, .., \vec{x}_L) \qquad (2.23)$$

or equivalently (going through the same derivation steps as above),

$$\max_{S_i,C_1,...,C_L} p(S_i) \prod_{j=1}^{L} p(C_j)p(\vec{x}_j|S_i,C_j) \qquad (2.24)$$

2.2.6 Paradigm 4: Incorporating Speaker-Specific Models Using *A Posteriori* Speaker Probability

Closer examination of the mathematical formulations derived thus far reveals that both Paradigms 2 and 3 make implicit use of the *a posteriori* probability for a given speaker over all available tokens, i.e. $p(S_k|\vec{x}_1,..\vec{x}_L)$. Paradigm 1, on the other hand, only makes use of the *a posteriori* probability by considering the tokens one at a time, i.e., $p(S_k|\vec{x}_i)$. Instead of using $p(S_k)$ in paradigm 1, we may be able to improve its performance by using $p(S_k|\vec{x}_1,..\vec{x}_L)$. Hope-

fully adjusting the *a-priori* probabilities of the speakers after looking at all the tokens would make one speaker more likely than others. As a result, the densities of that speaker would make a greater contribution in the classification process than in Eq. (1.12). This means that in the extreme case, when $p(S_k|\vec{x}_1, ..\vec{x}_L)$ is 1 for a certain speaker and 0 for all others, we use only the speaker specific-distributions for that speaker in making the decisions. Paradigms 3 and 4 are equivalent in that case.

Paradigm 1 uses speaker-specific models but imposes no constraints. Paradigms 2, 3 and 4 not only us speaker-specific models but also impose constraints in different ways. It might be worthwhile to keep in mind that there is an absolute baseline which is the most simple classification paradigm (which we call Paradigm 0).

2.2.7 Paradigm 0: Simple Bayesian Classification Using Pooled Data From All Speakers

In this case we do not distinguish between the different kinds of speakers there are. We simply collect them from all speakers, pool them together and use them to train the parameters to estimate $p(\vec{x}|w_i)$ for the training tokens. Our decision rule is the same as Eq. (2.8) and is rewritten here for convenience.

$$\max_{C_1, ..., C_L} \prod_j \frac{p(C_j)p(\vec{x}|C_j)}{p(\vec{x}_j)} \qquad (2.25)$$

2.3 A Toy Example

Before proceeding to experiments with real data, we conducted a very simple toy example to see the difference between Paradigms 1 and 3 under ideal model assumptions. The situation is similar to Figure 2.3 only with two speakers and two classes instead of three speakers and two classes. The

observation vector $\vec{x}$ is one-dimensional and has a Gaussian distribution. The notation is the same as developed in our mathematical formulation earlier. $N(m, \sigma^2)$ indicates a Normal distribution with mean m and variance σ^2.

$$p(x|S_1, w_1) = N(m1, 1)$$
$$p(x|S_1, w_2) = N(m2, 1)$$
$$p(x|S_2, w_1) = N(f1, 1)$$
$$p(x|S_2, w_2) = N(f2, 1)$$

Each test situation involved either Speaker 1 or Speaker 2 producing a sequence of L observation tokens. We compared results using Paradigms 1 and 3 in order to observe the difference between a speaker constraining paradigm and Paradigm 1.

2.3.1 Case I

$$m1 = 0, m2 = 4.0, f1 = 0.8, f2 = 5.0$$
$$p(w_1) = 0.3, p(w_2) = 0.7$$
$$p(S_1) = p(S_2) = 0.5$$
$$N = 2; n = 2; L = 200$$

We went through 65 sets of 200 tokens each produced by Speaker 1 and then another 65 sets of 200 tokens produced by Speaker 2. Performance of Paradigm 1 was 97.1% and performance of Paradigm 3 was 98.1%. The difference is significant at the 0.00001 level.

2.3.2 Case II

This time we moved the second speaker further away from the first in observation space thus increasing the difference between them.

$$m1 = 0, m2 = 4.0, f1 = 1.6, f2 = 6.0$$

$$p(w_1) = 0.3, p(w_2) = 0.7$$
$$p(S_1) = p(S_2) = 0.5$$
$$N = 2; n = 2; L = 200$$

Again we repeated the same number of experiments as in the previous case. This time performance of Paradigm 1 was 95.1% and that of Paradigm 3 was 98.6% (this difference is again significant at 0.00001). Clearly the difference between them seems to have increased as the speakers have moved apart. If the two speakers had identical characteristics, there would not have been any difference at all.

This simple example illustrates that there is potential room for improvement by imposing speaker constraints. Furthermore we have already seen that our attempts to break down overall distributions into speaker-specific ones might not be overly simplistic. The next chapter describes specific implementations on a real task.

2.4 Remarks

It is noteworthy that we have not at this point specified what the classes w_i refer to. They are linguistic units and could be words, phonemes, syllables or any other linguistic segments we choose as long as there are a finite, well-defined number of them. In our actual experiments we use phonemes as the recognition units. The preliminary correlation studies have also been done with phonemes.

The acoustic vector $\vec{x}$ produced when the speaker utters class w_i has been assumed to be a constant length vector. In reality, there is time variance in the speech signal and the length of a segment corresponding to a particular class will differ across speakers and across different realizations within the same speaker. Obviously some engineering approximation will have to be

used to time-normalize it. Furthermore, the actual acoustic representation of the speech signal is also left open. The theoretical framework requires one to be able to estimate $p(\vec{x}|w_i)$. How one does it is not explicitly dealt with in this thesis.

Similarly we have assumed there are N speakers or speaker types. How one obtains these speaker groups is unclear and is an open question. The four paradigms of recognition impose constraints in different ways to different degrees and have different computational requirements. From the toy example, it seems that the more separated the speaker groups are, the greater the difference in performance between speaker constraining paradigms (2, 3, and 4 although we tested only for 3) and Paradigm 1. Various other engineering issues come up in actually implementing them on a real task. The different sound classes and how they are distributed in acoustic space also has a bearing on the relative performance. These issues are raised and resolved in the next chapter on a specific task of vowel classification on the TIMIT corpus. This will give us an idea of the various tradeoffs involved among the paradigms, and we will be better able to compare and contrast them with each other.

Chapter 3

Comparison of Speaker-Constraining Recognition Paradigms on a Task of Vowel Classification

3.1 Introduction

In the previous chapter we formalized mathematically several different ways to enforce speaker constraints for the task of speech recognition. As mentioned in our remarks at the end, we left several things unspecified, such as what the pattern classes w_i refer to, the number of speaker types N, and ways to obtain them. Furthermore, our toy example seems to suggest that under ideal model conditions at least, it is meaningful to enforce speaker constraints. In this chapter we will describe several different experiments conducted on the task of vowel classification on tokens excised from the TIMIT corpus. This will help us evaluate the performance of our methods

of imposing speaker constraints on data collected from real speech.

We have described four different paradigms of recognition. All these paradigms decompose the overall population of speakers into several different speaker types. Paradigm 1 then assumes complete independence between different tokens produced by the same speaker. Paradigms 2, 3, and 4, on the other hand, impose constraints to different degrees in different ways. There are various engineering details which have to be taken care of in the implementation. We suspect that speaker-constraining paradigms (2, 3, and 4) would outperform Paradigm 1. We do not know, however, how much the difference would be, and whether it would be statistically significant. We also do not know how Paradigms 2, 3, and 4 compare amongst themselves. Besides, there is also Paradigm 0 which has no speaker models at all. Though it might be unfair to compare such a model with speaker constraining models, we would nevertheless do so from time to time since it is the prevailing method used in the speech recognition community. There are various other implementation issues which are likely to affect the relative performance of these paradigms. Some of them are:

- Task: The performance is going to depend on the task. If we are doing phoneme classification, the way in which speaker variability manifests itself might be different for different phonemes.

- Training Data: We have to estimate the parameters of our speaker-specific distributions $p(\vec{x}|w_i, S_j)$. Our estimates will depend on the amount of training data we have and this is going to affect performance. Some paradigms might be more robust than others.

- Number of Tokens We Optimize Over (L) : If $L = 1$, then speaker constraints are not really being applied at all, and the tokens are being treated independently. At this point we have no idea how large L must

40

be to meaningfully enforce speaker constraints. With increasing L our speaker constraining paradigms get more computationally expensive, which may become a concern.

- Speaker Groups: In our mathematical formulation we have assumed that the population consists of N speakers or speaker types. How one partitions the population into speaker types, and maintains a balance between capturing speaker variability through a large N and accurate estimates of the speaker's parameters based on limited training data is an open question.

- Representation of the Speech Signal: There are many ways to represent the speech signal. Some might capture speaker variability better while others might capture phonetic variability better. The trade-off between the two is also an issue and might affect the performance of the different paradigms.

- Classifier: We have formulated our problem in a classical Bayesian sense. However, the exact form of our densities is left open. Gaussian models might or might not fit the data closely. As we shall see later in this chapter, multi-layer perceptrons can be used to coerce *a-posteriori* probabilities from the data. We will look into the applicability of speaker-specific paradigms to phonetic classification using such a classifier.

- Computational Complexity: As we have mentioned, the paradigms differ in implementation and computational complexity. This might be of concern to us and might affect our choice of which paradigm to use.

This chapter probes at some of the above issues and attempts to get a better understanding based on empirical evidence. This will indicate the

feasibility of imposing speaker constraints for improvement in recognition performance for a certain task. What follows is a description of the experimental set-up, a roadmap of the experiments to be conducted, and an account of the experiments themselves. At the end of it all, we will hopefully have answers to some of the questions raised.

3.2 Task and Corpus

The corpus used was TIMIT, a description of which has been provided in Chapter 2. Our task was the classification of the eight vowels in American English, /i, ɪ, e, ɛ, æ, a, ʌ, ɔ/, using tokens excised from the above corpus. These eight vowels were chosen because a sufficient number of tokens of them are available for a set of test speakers, thereby enabling us to conduct valid experiments. Furthermore, the above set contains back and front vowels, and high and low vowels, and is thus representative of the different vowel types. Most of our detailed experiments are conducted on this smaller task to facilitate meaningful comparisons. In the next chapter, we report a few experiments on a larger task.

We selected 325 speakers who were designated as training speakers. There were 112 females and 213 males. Only the SX and SI sentences were taken, and all examples of the vowel tokens were extracted with no restriction placed on the phonetic environment of the extracted vowel tokens. Since the SX and SI sentences are different for different speakers, the phonetic environment varied from speaker to speaker. The actual procedure for this and the resulting representation of each vowel token is described in Section 3.3. There were 16324 training tokens in all.

65 speakers were selected as test speakers, out of whom 52 were male and 13 were female. The test speakers all had at least 4 tokens of each

	Number of Speakers (M/F)	Number of Tokens
Training	325 (213/112)	16324
Test	65 (52/13)	3670

Table 3.1: Corpus used for the experiments.

vowel class. In our theoretical formulation, we assumed $p(w_i|S_j) = p(w_i)$. We wanted to select test speakers in such a way that this assumption was not grossly violated. More importantly, our intent is to reduce confusions between similar vowels within a speaker by imposing speaker constraints. This could be more effectively achieved if there were a sufficient number of test tokens per vowel. The 65 test speakers yielded 3670 vowel tokens in all. The size and contents of the corpus are summarized in Table 3.1.

3.3 Signal Processing

The speech signal is sampled at 16 kHz and a spectral vector is computed every 5 ms. The 40-dimensional spectral vector is the output of an auditory model developed by Seneff [24], which will be described briefly.

3.3.1 Seneff's Auditory Model

Seneff's Auditory Model (SAM) has three stages, as illustrated in Figure 3.1. Stage I consists of a bank of 40 critical-band filters, spaced linearly on a Bark frequency scale. The center frequencies of these filters range from 130 to 6400 Hz. The outputs of this stage are fed into Stage II, which models the transformation from the basilar membrane vibration to the auditory-nerve fiber responses. This part of the model incorporates non-linearities such as dynamic range compression, half-wave rectification, short-term and rapid adaptation, and forward masking. The output of this stage represents
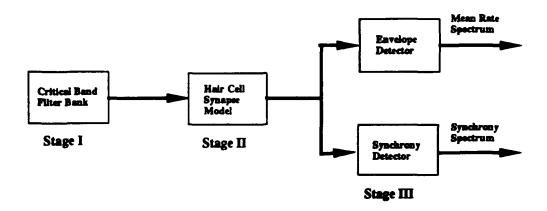
Figure 3.1: Seneff's Auditory Model

a profile of the probability of firing of the auditory-nerve fibers. This is processed by the envelope detector in Stage III to yield the mean probability of firing along the auditory nerve, called the mean rate response. The other module, the synchrony detector, determines the synchronous response of each filter by measuring the extent of dominance of information at the filter's characteristic response. Both the mean rate and the synchronous responses result in a 40-dimensional feature vector. In our experiments we used only the mean-rate response, and thus had one 40-dimensional vector per frame.

3.3.2 Time Normalization and Data Reduction

The different tokens excised from the different sentences all vary in duration, and hence there are a varying number of frames in their spectral representation. This presents a minor problem since we would like to have the vector $\vec{x}$ (in our mathematical treatment) to have the same dimension for all tokens. Time normalization is accomplished by taking the spectral average of the frames which constitute the middle-third of the vowel token, thus producing one 40-dimensional vector for each token. As a result, we had approximately 16000 data points of 40 dimensions in our training set.

We then did some dimensionality reduction by multiple discriminant

44

analysis[1] using Fisher's approach [4]. Multiple discriminant analysis is a way to project a d-dimensional space to a $c - 1$ dimensional space for a c class problem. Parametric or nonparametric techniques that might not have been feasible in the original space may work well in this lower-dimensional space. In particular, it may be possible to estimate separate covariance matrices for each class and use the general multivariate normal assumption after the transformation. For our eight vowel problem, we reduced the dimensionality from 40 to 7.

3.4 Model Assumptions and Implementation

Our mathematical framework defines the densities $p(\vec{x}|w_i, S_j)$. We have assumed in our implementation that these are Gaussian with a diagonal covariance matrix. Furthermore our covariance matrices and means are different depending on both speaker group and class. We assume that the *a-priori* probabilities of the occurrence of the different classes (vowels) are speaker-independent and known.

The implementations of these paradigms was done on a SUN SPARCstation in an S-Plus [1] software environment. S-Plus is a C-like language with a lot of functions for statistical analysis. It is also possible to write C-routines and call them from S-Plus. The latter has been done on occasions where the C-routines would be considerably faster.

---

[1]There are other ways to reduce dimensionality of data, the most common amongst them being principal components analysis [12]. This is applied in a later set of experiments in order to compare and contrast relative performance among the different recognition paradigms.

3.5 Roadmap of Experiments

We will now describe a series of experiments which were conducted on the above-mentioned task. These experiments were conducted in a controlled fashion in an attempt to resolve the issues we had raised in Section 3.1. For clarity of presentation, we have grouped these experiments into three categories on the basis of their broad similarities and differences. These are:

- **Experiment Set A:** In all the experiments belonging to this group, we perform supervised clustering of our speakers into male and female speaker groups. With this as a common feature, experiments have been conducted to investigate different representations of the vowel tokens, the influence of training set size, and the number of test tokens we optimize jointly (L).

- **Experiment Set B:** In this set of experiments, we chose our speaker groups by unsupervised clustering of the training speakers into N clusters. We investigated various clustering schemes by changing the clustering algorithms, and the representative vector space. Experiments which examine the influence of N and training set size on the relative performance of our recognition paradigms were also conducted.

- **Experiment Set C:** This consists of those experiments which can not justifiably belong to either of the sets above. Specifically, these experiments investigate issues which are relevant to experiments of both **A** and **B**, including computational complexity and the kind of classifier we use.

As we proceed through these experiments, we will comment on the trends observed, the control parameters altered and issues resolved.

3.6 Experiment Set A: Supervised Clustering

In this set of experiments we investigate some of the earlier issues but with a very specific way of choosing our speaker groups. We divide our speakers into two groups - male and female. This corresponds to supervised clustering, with $N = 2$. Given the nature of the anatomical difference between the vocal apparatus of males and females, there is reason to believe that such a gender grouping is reasonable.

3.6.1 Separation of Males and Females in Acoustic Space

Shown in Figure 3.2 are the male and female centroids, i.e., the estimated means of the gender-specific probability distributions for the different vowels, in the space spanned by the first and second discriminant functions. There are a few interesting observations we could make here. The male and female centroids are different indicating that males and females have different acoustic characteristics. Further, it appears that the male acoustic space is rotated and shifted to give the female acoustic space. This is clearer when one takes only front vowels and performs a linear discriminant analysis on them as shown in Figure 3.3 or if one takes only the back vowels and performs one on them separately.

In order to assess if the apparent difference between male and female centroids is statistically significant, we conducted a simple test. For each vowel, we took the male and female populations and tested $H_0 : \mu_m = \mu_f$ against $H_1 : \mu_m \neq \mu_f$ where μ_m is the male mean and μ_f is the female mean for that vowel class. In each case (i.e. for each vowel class) the null
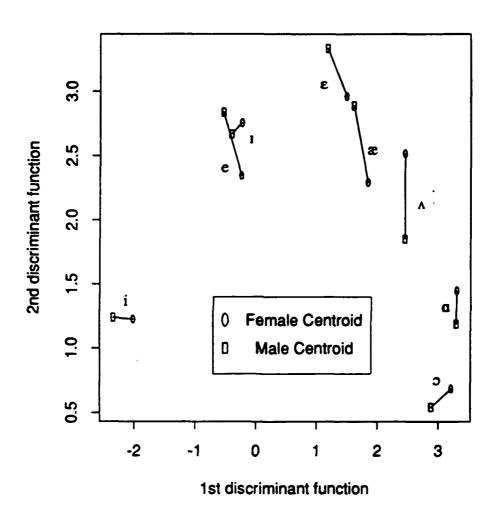
Figure 3.2: Comparison of the male and female centroids displayed in the space spanned by the first and second discriminant functions.
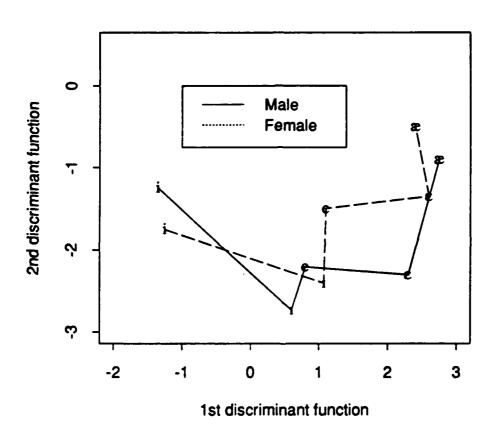
Figure 3.3: Male and Female centroids with a linear discriminant analysis done on front vowels only. The centroids have been connected together to show how the space is rotated.

hypothesis was rejected with very low p-values ($p < 0.001$), thus indicating that the male and female clusters are well separated.

3.6.2 Training Set Size

Shown in Figure 3.4 are the relative performances of Paradigms 1 through 4 on both test and training data, as a function of the amount of training data used to estimate the parameters of our models. At full training set size, Paradigm 1 performs at 60.27% recognition accuracy. Paradigms 2 and 3 have identical performance at 61.69% indicating an improvement of 1.4% over the baseline. This difference in performance is significant at the 0.005 level using McNemar's Test [7]. Paradigm 4 yields the best performance at 61.93%, an improvement of 1.7% over the baseline (again statistically significant, this time at the 0.001 level). This is very satisfying because it indicates that by employing speaker consistency at a primitive level, i.e., employing gender consistency, we have managed to get significant improvement. When we performed the same experiment but with a smaller fraction of the training data, we randomly drew a fraction of the training tokens for each vowel class while maintaining the male-female ratio. We performed experiments at 2, 5, 11, 17, 20, 40, 60, and, 80% training set sizes. At each training set size, we repeated the experiment approximately 7 times. More repetitions were performed at small training set sizes and fewer were performed at larger training sizes. Since we were randomly picking a fixed fraction of the tokens, we got several different classification accuracies for each paradigm at each size. What is plotted in the figure is a smoothed version of this raw data to show the general trend.

There are certain other interesting observations. The performance on test data for each paradigm increases with training set size. This is reasonable as estimates of model parameters improve with more training data. At the same

50
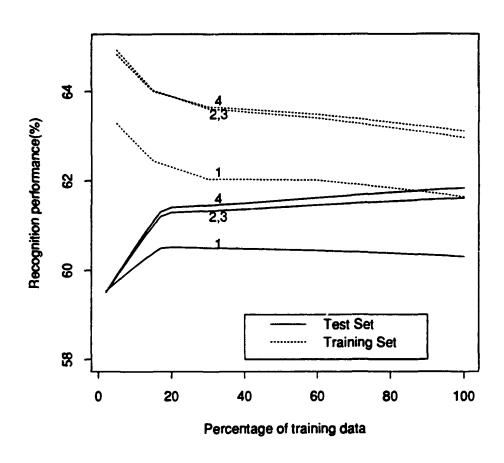
Figure 3.4: Vowel classification performance on training and test data for the four paradigms, plotted as a function of the amount of training data.

time the difference between the performance on test data and that on training data decreases indicating that our models and our estimates generalize well with increasing training data. It is noteworthy that performance using Paradigm 1 seems to have reached an asymptote while that for the other paradigms still seem to be improving. For large training set sizes ($\geq 20\%$ of full training data) there is a difference in performance between Paradigms 1 and 2, 3, and 4 with the speaker (gender, in this case) constraining paradigms having a higher classification accuracy. This difference is significant at the 0.01 level using McNemar's Test. However, there is very little difference between the different speaker constraining paradigms. Paradigm 4 seems to perform slightly better, but this difference is not significant even at the 0.05 level. For smaller training set sizes, there is very little difference between the performance of the different paradigms. This could be due to the fact that at lower training set sizes, we have poorer estimates of the male and female parameters. As a result, forcing the gender constraint using these poorly estimated parameters is not necessarily useful. As a matter of fact, when we tested to see if there was a difference between male and female means for small training set sizes, we often found that the significance level had increased indicating that males and females were not necessarily well separated any more. The same trends are roughly observed when we tested on the training data. For the record, Paradigm 0 performed at 60.1% classification accuracy at full training set size. Its performance was consistently poorer than Paradigm 1 for smaller training set sizes as well. However, the difference ranged from 0.2% to 0.5% and was not significant.

Another very interesting observation is that Paradigms 2 and 3 yield identical results in almost every single experiment. There is again a reason

for this. Recall that Paradigm 2 performed the following optimization:

$$\max_{C_1,...,C_L} \sum_{i=1}^{N} p(S_i) \prod_{j=1}^{L} p(C_j)p(\vec{x}_j|S_i,C_j) \tag{3.1}$$

We choose the C_i's to optimize a sum of different product terms. In our case $N = 2$, and it turns out that one of the product terms completely dominates the other. As a result maximizing the sum of these two disparate terms is equivalent to just maximizing the larger of the two. But maximizing the larger of the two is exactly what Paradigm 3 does and hence the two results are identical.

Shown in Table 3.2 are confusion matrices of the kinds of errors made by Paradigms 1 and 3. Confusions between similar vowels, such as /ɑ/-/ɔ/ and /e/-/ɪ/ have decreased in Paradigm 3. It is our belief that since the speech articulators are in similar positions for similar sounds, these similar sounds are more likely to be correlated. Hence, imposing speaker constraints will exploit these correlations and reduce confusions between these sounds.

3.6.3   Number of Test Tokens Jointly Optimized at a Time (L)

As has been mentioned befoɪ ᵔ, we impose speaker constraints by classifying tokens jointly (L at a time). According to our theoretical formulation, when $L = 1$, we effectively impose no speaker constraints at all. With increase in the value of L, the degree of speaker constraints increases.

To investigate the behavior of the recognition paradigms with change in L, we took all our training data and estimated speaker-specific distributions as before. For our test tokens, we took each speaker and collected his or her tokens. We decided on a value of L (which was maintained for all speakers) and randomly divided the speaker's tokens into groups of L tokens each.

Paradigm 1

	ɑ	e	i	ɛ	ʌ	ɪ	ɔ	æ
ɑ	70.3	0	0	0.84	11.2	0	13.7	3.92
e	0	58.8	10.2	9.12	0.55	16	0	5.25
i	0	7.65	82.1	0.14	0.29	9.38	0	0.43
ɛ	2.15	14.2	0.43	40.1	10.3	11.4	1.29	20.2
ʌ	17.2	1.23	0	14.5	43.6	4.9	11	7.6
ɪ	0	20.3	10.4	9.62	4.43	52.4	0.76	2.14
ɔ	25.3	1.15	0	0.57	4.31	2.59	66.1	0
æ	5.25	11.6	1.05	13.9	3.94	0.52	0.52	63.2

Paradigm 3

	ɑ	e	i	ɛ	ʌ	ɪ	ɔ	æ
ɑ	72.8	0	0	0.28	11.5	0	11.5	3.92
e	0	60.8	10.2	9.67	0.55	13.5	0	5.25
i	0	7.65	81.8	0	0.29	9.81	0	0.43
ɛ	2.58	14.4	0.64	41.9	11.2	9.87	1.29	18.2
ʌ	18.6	0.98	0	14.2	47.3	3.68	8.09	7.11
ɪ	0	19.4	9.47	10.2	5.04	53.3	0.46	2.14
ɔ	26.4	0.86	0	1.15	4.6	1.72	65.2	0
æ	4.99	9.71	1.05	13.1	3.94	0.26	0.52	66.4

Table 3.2: Confusion matrices of Paradigms 1 and 3 on vowel classification task at full training set size. Speaker groups were based on gender.
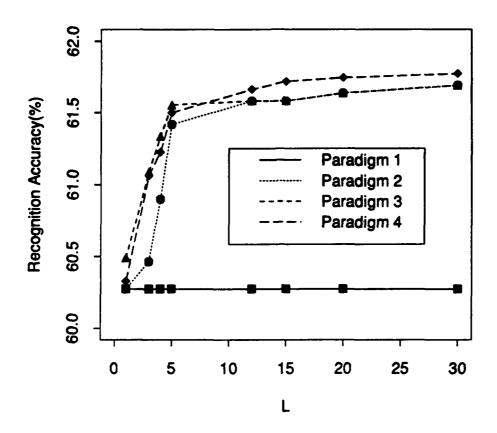
Figure 3.5: Variation of recognition accuracy with L.

Since all the speakers did not have exactly the same number of tokens, this grouping could not be exact and we often had one group which contained the left-over tokens. In any case, these groups of tokens were then classified using Paradigms 1 through 4. Furthermore, since grouping into tokens was random, we did the experiment several times for each value of L. The experiment was repeated for values of $L = 1, 3, 4, 5, 12, 15, 20, 30$. The results are shown in Figure 3.5.

The results are again as predicted. Paradigm 1 assumes independence between tokens and is independent of L. This exactly what is observed in our

experiments. As the value of L decreases, the difference between the speaker constraining paradigms and Paradigm 1 decreases. As expected, Paradigms 1 and 2 yield identical results for $L = 1$, since the equations used become equivalent. Paradigms 3 and 4 have slightly different equations for the $L = 1$ case. Hence their performance is slightly different, though comparable. It is interesting to note that for low values of L, Paradigms 2 and 3 have different results. For higher values, however, the same dominance of one term starts to take over and we have identical results again. For the record, at $L = 1$, Paradigms 1 and 2 yield 60.27% accuracy while Paradigm 3 yields 60.49% and Paradigm 4 yields 60.33%. None of these are significantly different from one other.

3.6.4   Principal Components Analysis

The experiments described above were conducted using linear discriminant functions as a technique for data reduction. This involved rotating the original 40-dimensional space in one particular way, and creating a particular form of representation. To see if the above trends are independent of representation, we decided to use principal components analysis [12] to reduce the dimensionality. Principal components analysis defines a rotation of the dimensions of $\bar{x}$. The first derived direction is chosen to maximize the standard deviation of the derived variable, the second to maximize the standard deviation among directions uncorrelated with the first, etc.

Shown in Figure 3.6 are the male and female centroids for each vowel class. This has been plotted in the space spanned by the first and the second principal component. The figure shows how the male space seems to be shifted to obtain the female space. The transformation from males to females seems to be much simpler in this case, as compared to that based on linear discriminant analysis. It is important to note that looking at these
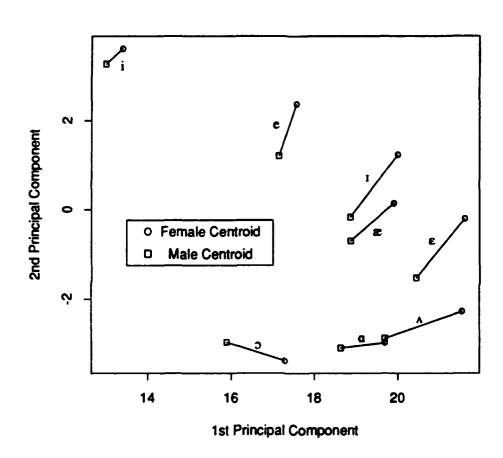
Figure 3.6: Comparison of the male and female centroids displayed in the space spanned by the first and second principal components.

kinds of figures might be misleading because whatever patterns emerge in the two dimensions which have been plotted might not be true when considering the entire vector space. Unlike linear discriminant analysis, principal components analysis yields a 40 dimensional vector. Principal components analysis achieves two useful objectives. Firstly, it diagonalizes the vector space so that the different dimensions are no longer correlated. This makes our diagonal covariance assumption more reasonable. Secondly, the dimensions are arranged in order of variance, i.e., the first component captures the most variance, the second dimension captures the second-most variance etc. If we use the top few principal components only, we would have achieved data reduction. However, how many components to use is an open question.

Shown in Figure 3.7 is the performance of Paradigm 0 with varying number of dimensions. Performance seems to have levelled off after 10 dimensions and in fact actually drops. Consequently we decided to conduct our detailed experiments with the first 12 *principal components* which captured approximately 96% of the variability. The reason we used Paradigm 0 to decide the number of components to use is that it is the absolute baseline, which makes no speaker assumptions whatsoever, and thus is not biased towards any of the other paradigms.

We used all the training data and obtained gender-specific distributions just as before. At full training set size, Paradigm 1 performed at 62.02% accuracy and Paradigms 2, 3, and 4 operated at 63.05% accuracy. The difference is significant at the 0.01 level using the McNemar's test. We also conducted several different experiments using 20, 40, 60, and 80% of the data. When using a fraction of the data, we picked tokens at random for each vowel while maintaining the male/female ratio just as before. Table 3.3 contains the average performance of the different paradigms for varying training set sizes.
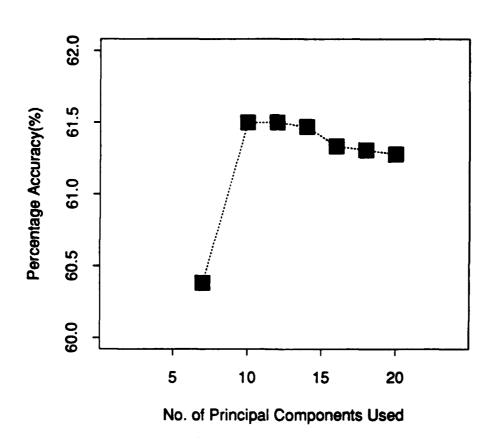
Figure 3.7: Variation of recognition accuracy with number of principal components used for Paradigm 0 at full training set size.

Training Data Used (%)	20	40	60	80	100
Paradigm 1 Accuracy (%)	61.66	61.96	61.91	62.10	62.02
Paradigms 2,3,4 Accuracy (%)	62.67	62.89	62.97	63.06	63.06

Table 3.3: Performance of the different paradigms as a function of training data with principal components analysis applied to reduce dimensionality.

We find again that the speaker constraining paradigms perform better than Paradigms 0 and 1. The difference is of the order of 1% which is less than before. It is, however, still statistically significant at the 0.01 level using McNemar's Test. This suggests that improvement in performance on applying speaker constraints is independent of the method used to·reduce dimensionality. It is also noteworthy that Paradigms 2, 3, and 4 provide identical results. The reasons for the identical performance of Paradigms 2 and 3 are the same as before. As for Paradigm 4, it turns out that $p(S_i|\vec{x_1},..\vec{x_L})$ is usually always 1 or 0 for each speaker[2]. This reduces it to Paradigm 3.

3.6.5 Representation of Vowel Tokens Using Three Slices.

Some of the above experiments investigated different representations of the vowel tokens but measurements were made only on the middle-third of each vowel. We also examined another representation of the vowel tokens. This time, we took each vowel token and divided it into three equal parts along its time axis. Then we obtained spectral averages for each part. Thus we had spectral averages for the first-third, middle-third and last-third of each token. Each vowel token was hence represented by three vectors of 40 dimensions

---

[2]The reason for this is somewhat unclear, but it could be because we use 12 principal components but only 7 discriminant functions. These get multiplied causing greater disparity in the $p(S_i)$'s.

each. For data reduction, we employed the technique of linear discriminant analysis. However, a different rotation was performed on each of the three parts of the vowel, resulting in three separate seven-dimensional vectors for each vowel. The overall feature vector for each vowel token was obtained by simply concatenating these three vectors together to yield a 21 dimensional vector. When we performed our classification using the different paradigms, we found that Paradigm 1 performed at 59.29% accuracy and Paradigm 3 performed at 60.30%. We did not perform the other paradigms because by now we were reasonably convinced that there was not a significant difference between the different speaker constraining paradigms, at least using gender-specific models. Though the difference between Paradigm 1 and 3 was significant at the 0.01 level, the absolute performance was rather low. We suspect this was due to the diagonal covariance assumption in our Gausssian classifier. Recall that our feature vector was a concatenation of three vectors representing three segments in time. The dimensions for each of those vectors are uncorrelated within themselves due to the nature of the linear discriminant analysis but they are correlated with the dimensions of other vectors. Thus, in our concatenated feature vector, $x[1]$ and $x[3]$ are uncorrelated but $x[1]$ and $x[8]$ may be correlated. To verify this hypothesis, we could either do away with the diagonal assumption and use a full covariance matrix, or transform our feature vector space using a principal components transformation. We chose to do the latte Shown in Figure 3.8 is a performance of Paradigm 1 and Paradigm 3 with varying number of dimensions used. Again observe that Paradigm 3 has a higher recognition accuracy than Paradigm 1 but the difference is not always significant at the 0.01 level. The reason for this is not clear. Note also that the absolute recognition accuracy has increased by about 5%.
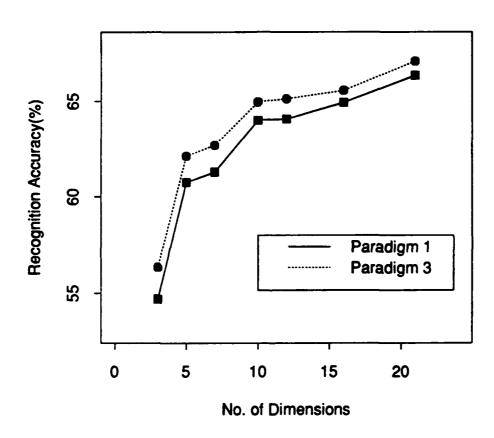
Figure 3.8: Variation of recognition accuracy with number of dimensions. Here the vowel is represented by spectral average of three slices. The data is diagonalized using principal components analysis as described in text.

3.6.6 Summary

All of the experiments of Set A used only gender-specific models. Some of the parameters varied were the training set size, number of tokens optimized at time (L) and representation and method of data reduction. The general conclusions could be reiterated as follows:

- Speaker constraining paradigms perform better than Paradigm 1 (and also Paradigm 0). The actual performance increase varies from 1-2% for our task depending upon the kind of representation used. At low training set sizes, this difference becomes insignificant. For high-dimensional feature vectors, this difference is somewhat smaller and often insignificant.

- The different speaker constraining paradigms do not differ significantly from one another. In a lot of cases using gender-specific models, they actually yield identical results.

- As the number of tokens we optimize over decreases, the difference between the speaker constraining paradigms and Paradigm 1 decreases and eventually becomes insignificant. In fact Paradigms 1 and 2 are mathematically equivalent in the $L = 1$ case.

3.7 Experiment Set B: Unsupervised Clustering

In this set of experiments we investigated alternate ways to group our speakers. Unsupervised clustering has been tackled quite often in the past, especially in the fields of Statistics and Pattern Recognition [4]. As we shall see, clustering speakers into meaningful groups is a very difficult task and no one

solution is clearly correct. There are many reasons for this. Firstly, we do not know how to characterize each speaker, i.e., how to get a feature vector which would contain information about the speaker's acoustic characteristics. The different test speakers have all produced different numbers of vowel tokens. How to combine them to obtain a vector of the same dimension for all speakers is an open question. Secondly, several different algorithms exist for clustering. Thirdly, we do not know how many clusters one should have. There is no well defined optimality criterion for this. There is a tradeoff between having enough clusters to capture the variability among the speakers and having enough speakers in each cluster to estimate the cluster-specific model parameters well.

3.7.1   Space in Which to Cluster the Speakers

As has been mentioned before, the different speakers have produced different numbers of tokens for each vowel class. We would like to utilize these tokens effectively, and produce a vector of fixed dimensions so that each speaker can now be characterized by this representative vector in the same acoustic space. One straightforward method would be to simply average all the tokens produced by each speaker without paying any heed to which class they belong. In this case letting $\vec{y_i}$ refer to the feature vector for the ith training speaker we have

$$Representative\ Vector\ 1 = \vec{y_i} = \frac{1}{L_{tri}} \sum_{j=1}^{j=L_{tri}} \vec{x}_{ij} \qquad (3.2)$$

Here L_{tri} is the total number of tokens produced by the ith training speaker and $\vec{x}_{ij}$ is the feature vector for the jth token produced by the ith training speaker. This feature vector could be in the reduced space spanned by the linear discriminant functions or the principal components of the hair-cell

representation of the middle-third of the vowel token. If a certain speaker has produced mostly front vowels and another has produced mostly back vowels, then, the vectors ($\vec{y}$) calculated for each of them are going to be quite different, although they might have very similar acoustic characteristics in general. Hence, the class to which each of the speaker's tokens belong must be taken into consideration. We investigated four different ways to combine a speaker's tokens. The first was the simple method shown in Eq. 3.2 The second was

$$Representative\ Vector\ 2 = \vec{y_i} = \frac{1}{L_{trifront}} \sum_{j=1}^{j=L_{trifront}} \vec{x}_{ijfront} \qquad (3.3)$$

where $\vec{x}_{ijfront}$ is the jth front vowel token produced by the ith speaker. There are $L_{trifront}$ front vowel tokens produced by the ith training speaker and these were all pooled together. The third was

$$Representative\ Vector\ 3 = \vec{y_i} = \frac{1}{L_{triback}} \sum_{j=1}^{j=L_{triback}} \vec{x}_{ijback} \qquad (3.4)$$

where $\vec{x}_{ijback}$ is the jth back vowel token produced by the ith training speaker and there are $L_{triback}$ back vowel tokens produced by that speaker in all. Finally, we also gave individual importance to each vowel class. We obtained our feature vector as follows:

$$Representative\ Vector\ 4 = \vec{y_i} = \frac{1}{8} \sum_{j=1}^{j=8} \vec{\bar{x}}_{ij} \qquad (3.5)$$

Here $\vec{\bar{x}}_{ij}$ represents the mean of the tokens belonging to the jth vowel class and produced by the ith speaker. We shall describe the exact experiments conducted with these clustering schemes later.

3.7.2　Algorithm Used to Cluster

In the previous section we talked about ways to obtain a representative vector for each speaker. In our case, we have 325 training speakers and hence 325 vectors in all which we would like to cluster into different groups. Two methods [4] were investigated:

Hierarchical Clustering

In the beginning there are N_s clusters, where N_s is the total number of speakers. In our case $N_s = 325$. At each stage, the "nearest" clusters are combined to form a bigger cluster. The distance between two clusters can be defined according to our will. In our experiments, we chose the largest Euclidean distance between points in one cluster and points in another cluster to be the distance between the two clusters. This avoids the formation of long thin clusters and tries to form more spherical clusters. Hierarchical clustering continues to aggregate groups together until there is just one big group. At every stage of combining two groups, a note of the distance metric is made. This distance metric is lowest for the first grouping (since the closest clusters are grouped) and highest for the last grouping. The clusters are formed in this fashion until only the desired number of clusters are left.

K-means

This is one of the more popular non-hierarchical methods used. Here we have again $N_s = 325$ points which are to be divided into K clusters. We start with K initial cluster centroids (seed points) which are picked at random from the N_s points. Then we proceed through the list of points, assigning each point to the cluster whose centroid is "nearest". In our experiments, we used a Euclidean distance metric. After this has been done for all points, we

66

recompute the cluster centroids and repeat the process again. This is done until no more reassignments take place.

3.7.3 Clustering Experiments

The purpose of these experiments was to determine the efficacy of different methods of clustering speakers. We started out with the 16324 training tokens from the 325 speakers. Each token was represented by a 40-dimensional vector which was the spectral average of the middle-third. Linear discriminant analysis was done as before to reduce the number of dimensions to 7. Then the representative vector for each speaker was computed using the four methods outlined in Section 3.7.1. We used both K-means and hierarchical clustering thus yielding 8 different methods of clustering. There really is no way of deciding which method of clustering is reasonable. It is our belief, however, that if one were to divide the speakers in the world into two clusters, one cluster would be predominantly male and the other would be predominantly female. Using this as our yardstick, we decided to cluster the speakers into two groups using various methods and observe how closely the clustering corresponds to the gender of the speakers. We show in Table 3.4, contingency tables indicating what percentage of the total speakers were in each cluster and respectively male and female.

It would be meaningful to measure the correlation between which group a speaker lies in, and his or her gender. In other words, how much information is provided by the speaker's group about the gender. This can easily be cast in information-theoretic terms and we can measure the mutual information between the two methods of clustering (supervised into gender groups and unsupervised into the two classes.)

It might help at this point to provide some background on entropy and mutual information [6]. Suppose we have a random variable X. The entropy

67

	K-means	
	Cluster 1	Cluster 2
Males	39.7	25.8
Females	14.5	20.0

Rep. Vector 1

	Hierarchical Clustering	
	Cluster 1	Cluster 2
Males	60.9	4.6
Females	33.8	0.7

	Cluster 1	Cluster 2
Males	57.8	7.7
Females	8.5	26.0

Rep. Vector 2

	Cluster 1	Cluster 2
Males	52.9	12.6
Females	30.8	3.7

	Cluster 1	Cluster 2
Males	47.3	18.2
Females	9.3	25.2

Rep. Vector 3

	Cluster 1	Cluster 2
Males	41.4	24.1
Females	27.2	7.3

	Cluster 1	Cluster 2
Males	41.4	24.1
Females	14.1	20.4

Rep. Vector 4

	Cluster 1	Cluster 2
Males	44.7	20.8
Females	24.0	10.5

Table 3.4: Clustering of speakers into two groups by different algorithms using different representative vectors. Dimensionality reduction is done by linear discriminant analysis.

of the random variable is defined as

$$H(X) = -\sum_{x} P_X(x) log(P_X(x)) \qquad (3.6)$$

This entropy is a measure of the average uncertainty in X. We can also define the conditional entropy of X after observing another random variable Y. Thus

$$H(X|Y) = -\sum_{xy} P_{XY}(x,y) log(P_{X|Y}(x|y)) \qquad (3.7)$$

In the above equations, $P_X(x)$ is the probability distribution of X, $P_{X|Y}(x|y)$ is the conditional probability distribution of X given Y and $P_{XY}(x,y)$ is the joint probability of X and Y. $H(X|Y)$ is thus the average uncertainty in X after observing Y. The mutual information $I(X;Y)$ between X and Y is defined as the average reduction of uncertainty in X after observing Y. It follows:

$$I(X;Y) = H(X) - H(X|Y) \qquad (3.8)$$

In our problem, we can imagine drawing a speaker out of the population and defining the random variable X to be 0 if the speaker is female and 1 if the speaker is male. We define the random variable Y to be 0 if the speaker belongs to Cluster 1 using the clustering scheme shown and 1 if the speaker belongs to Cluster 2 using the same clustering scheme. The mutual information between the two variables would be high if there was a close correlation. If the two variables were completely independent, then the mutual information would be 0. We estimate the distributions from our tables and Table 3.5 shows the mutual information in each of the cases.

We also looked into an alternative representation for clustering. We reduced our 40-dimensional space using principal components analysis as described previously. We then took the first twenty principal components so

K-means		Hier. Cluster.
0.023	*Rep. Vector 1*	0.009
0.300	*Rep. Vector 2*	0.009
0.139	*Rep. Vector 3*	0.019
0.032	*Rep. Vector 4*	0.000

Table 3.5: Mutual Information between unsupervised clusters and gender. Dimensionality is reduced using linear discriminant analysis.

that now each vowel token was represented by a 20-dimensional vector. We repeated the same experiments as before. Table 3.6 shows the correlation of gender with speaker grouping for the various cases and Table 3.7 shows the the mutual information for the clusters formed in these cases.

It is worthwhile to note that there is a strong similarity in the trends observed in the two cases. For some reason which is not clear, the K-means method yields clusters which are better correlated to gender than the hierarchical clustering method. This is observed regardless of the method used to obtain the representative vector for each speaker. Furthermore, for a K-means clustering scheme, using *Representative Vector 1*, (i.e averaging all tokens without regard to class for each speaker) seems to do the worst in clustering speakers into gender classes. This is not surprising as this representative vector is highly dependent on the number of tokens belonging to each class produced by the speaker and this is not the same from speaker to speaker. When we take the average of front vowels only, i.e, *Representative Vector 2*, we get the best separation. Finally, taking the first 20 principal components we get better separation than using the linear discriminant function representation as evidenced by correspondingly higher mutual information values. We have plotted in Figure 3.9 a scatterplot of how the different

K-means   **Hierarchical Clustering**

	Cluster 1	Cluster 2
Males	23.4	42.1
Females	25.9	8.6

Rep. Vector 1

	Cluster 1	Cluster 2
Males	21.0	44.5
Females	22.5	12.0

	Cluster 1	Cluster 2
Males	59.3	6.2
Females	6.7	27.8

Rep. Vector 2

	Cluster 1	Cluster 2
Males	38.0	27.5
Females	30.8	3.7

	Cluster 1	Cluster 2
Males	16.7	48.8
Females	25.6	8.9

Rep. Vector 3

	Cluster 1	Cluster 2
Males	62.3	3.2
Females	33.6	0.9

	Cluster 1	Cluster 2
Males	44.8	20.7
Females	8.9	25.6

Rep. Vector 4

	Cluster 1	Cluster 2
Males	30.0	35.5
Females	8.3	26.2

Table 3.6: Clustering of speakers using different algorithms and different representative vectors. Dimensionality reduction is done using principal components analysis.

K-means		Hier. Cluster.
0.105	*Rep. Vector 1*	0.073
0.384	*Rep. Vector 2*	0.083
0.162	*Rep. Vector 3*	0.000
0.122	*Rep. Vector 4*	0.034

Table 3.7: Mutual Information between unsupervised clusters and gender. The dimensionality is reduced using principal components analysis.
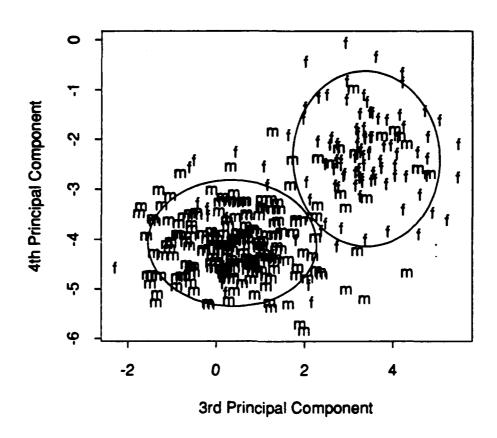
Figure 3.9: Distributions of the speakers in the space spanned by the 3rd and 4th dimension of the speaker's representative vector.

speakers are distributed in the space spanned by *Representative Vector 2* using a principal components representation for the individual vowel tokens. The separation of the speakers into gender classes is best seen in the third and fourth dimension. Also plotted are confidence ellipses for the two clusters obtained using K-means. Each ellipse includes 95% of the speakers belonging to each cluster. It can be seen that by and large, the males are included in one cluster and the females are included in the other.

3.7.4 Variation of Performance of the Different Recognition Paradigms with Number of Speaker Groups (N)

We have thus far only addressed the issue of the methods used to cluster speakers. The other question we have not resolved yet is how many clusters we should have, i.e., what is the value of N in our mathematical formulation. Again there isn't a single obvious way to decide what the optimal N is. A better question to ask may be how the different paradigms compare with each other with changing values of N both on a relative (i.e., in terms of which paradigm is the best and which is the worst) and an absolute scale (i.e., what actual percentage accuracy is achieved by each).

The first set of experiments we conducted towards this end is as follows. Our training set was still the same 16324 tokens and we reduced the space using linear discriminant analysis as usual. We divided the speakers into N groups using K-means and using *Representative Vector 1* for each speaker. Recall that this was the most basic clustering scheme and actually gave poorest separation into gender classes for $K = 2$. We estimated speaker-group-specific distributions, $p(\vec{x}|w_i, S_j)$ and assumed these were Gaussian with diagonal covariance matrices. Paradigms 1, 2, 3, and 4 were implemented as usual and shown in Figure 3.10 are their classification accuracies for $N = 2, 4, 8, 16$, and 32.

Paradigms 2, 3, and 4, true to previous results, do not show any significant difference in performance with respect to each other. All of them however perform significantly better than Paradigm 1 for 4 and 8 clusters. Interestingly, Paradigm 1 gradually increases in performance with increase in the number of clusters, while the other paradigms show a distinct peak with optimal performance around 8 clusters. It is our belief that as the number of
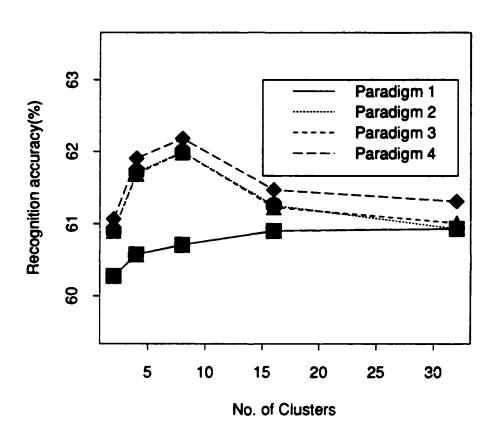
Figure 3.10: Variation of recognition accuracy with number of clusters at full training size. The clusters are obtained by K-means using *Representative Vector 1* for each speaker. The data was reduced using linear discriminant analysis.

clusters becomes very large, the estimates of the cluster-specific parameters become poor due to sparse data problems. Consequently, forcing these cluster constraints actually yield diminished performance. In Paradigm 1, there is no forcing of such cluster constraints, and hence there is no decrease in performance. In the other paradigms, however, such a decrease is observed at large values of N. Furthermore, we suspect that $N = 8$ is optimal only for this training set. If the training set were to increase in size, presumably it would take much larger values of N for the poor estimation problem to start manifesting itself. If it were to decrease in size, the reverse would be true. To investigate this issue, we conducted an experiment on the same task, this time using the training tokens from only 248 speakers instead.

The data was again reduced by using linear discriminant analysis and the recognition was done for the same number of clusters as before. Figure 3.11 shows the difference in performance between the paradigms for this reduced data set. There are two observations to be made here. First, for the same number of clusters, using all the training data gives higher performance. This is not surprising, since more training data gives better estimation of cluster-specific parameters. Secondly, and more interestingly, with less training data, the peak for Paradigms 2, 3, and 4 has shifted back to around $N = 4$. This seems to indicate that with less training data, reasonable estimates of cluster-specific parameters occurs for lower values of N. Thus the values of N for which it is profitable to impose the speaker constraint have decreased. So while the drop had started only after $N = 8$ in the earlier experiment, here the drop starts after $N = 4$.

These findings are generally indicative of the trade-off between the need for a sufficient number of clusters and the availability of training data. The above experiments suggest that as N increases, the speaker variability is better captured. Hence the speaker constraining paradigms work better, and
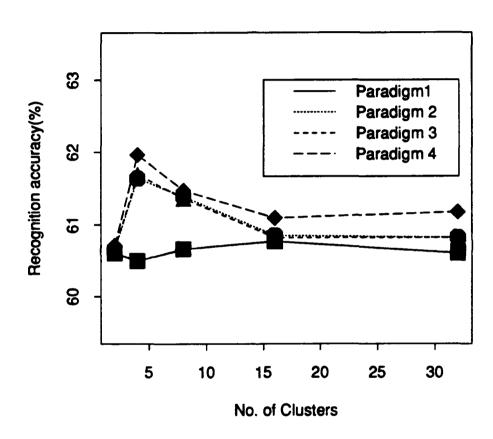
Figure 3.11: Variation of recognition accuracy with number of clusters at 75% training size. The clusters are obtained by K-means using *Representative Vector 1* for each speaker.

their recognition performance increases. However, as N increases, there are also fewer speakers per cluster and therefore poorer estimation of cluster-trained parameters. After a point, this phenomenon catches up and we start to observe a decrease in performance. Where the peak in performance occurs depends upon the total amount of training data used.

In our second set of experiments, the only thing we changed was the method we used to obtain our speaker clusters. Instead of using the most intuitive albeit the worst possible method, (i.e., using the mean of all vowel tokens as a representative vector and linear discriminant analysis for data reduction), we now used the best possible method (i.e. using the mean of only the front vowel tokens as a representative vector with each vowel token represented by its first 20 principal components). Figures 3.12-14 show how this method works using all the training data, then only 248 speakers out of the 325, and then finally 125 out of the 325 speakers. The trends observed are the same. For full training set size, speaker constraining paradigms show a peak at $N = 4$ with the speaker constraining paradigms performing significantly better than Paradigm 1 at $N = 2, 4, 8$.

When we use only 248 speakers, this peak is somewhat flattened out and the difference between the cases with 2 and 4 clusters is only very slight. Finally, when we use only 125 speakers, we find that this fall in performance is very dramatic, and the performance of speaker constraining paradigms decreases consistently from the $N = 2$ case with disastrous performance at high values of N. For this situation, there is no difference between Paradigms 1, 2, 3, and 4 for low values of N. At high values, Paradigm 1 does significantly better.
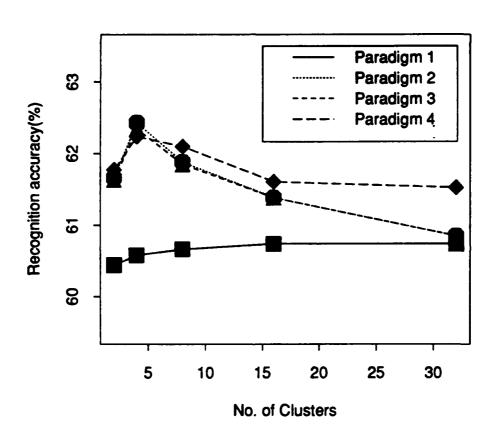
Figure 3.12: Variation of recognition accuracy with number of clusters at full training set size. Clusters are obtained using K-means and *Representative Vector 2* in principal components' space.

Figure 3.13: Variation of recognition accuracy with number of clusters with 248 speakers. Clusters are obtained using K-means and *Representative Vector 2* in principal components' space

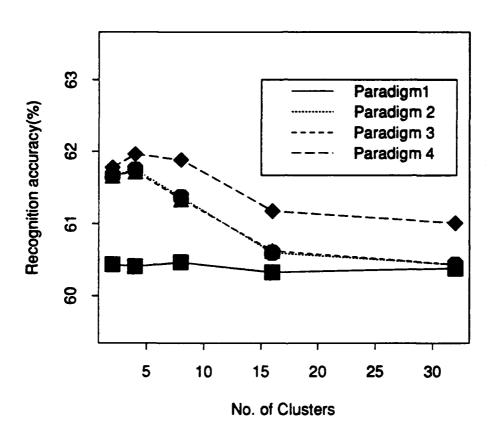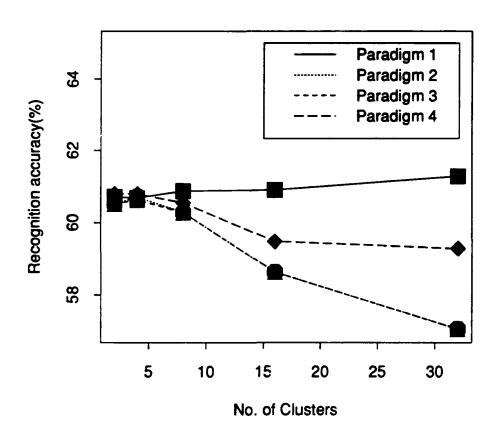Figure 3.14: Variation of recognition accuracy with number of clusters with 125 speakers. Clusters are obtained using *K*-means and *Representative Vector 2* in principal components' space.

3.7.5 Summary

In this set of experiments we investigated unsupervised techniques to group our training speakers into representative clusters. Furthermore, we examined the influence of N, i.e., the number of such clusters on the relative performance of Paradigms 1 through 4. The major conclusions are iterated again:

- We investigated four methods of combining the different tokens produced by the test speaker. We found that representing each speaker by the average of his/her front vowels provided best separation into male and female groups.

- In similar vein, we found that representing each speaker's tokens by the first 20 principal components was superior to representing each speaker's tokens by the 7 discriminant functions. Furthermore, K-i. · ; ielded better clusters than hierarchical clustering.

- On varying N, we found that Paradigm 1 showed steady improvement in performance. However, the other paradigms yielded optimal performance for some values of N and lower performances for very small or very large N. The optimal value of N depended upon the amount of training data used and the nature of the feature vector.

3.8 Experiment Set C: Other Related Experiments

In the above set of experiments we dealt with issues of training size, forming speaker groups and different representations. In this section, we explore the issues of computational complexity and the kind of classifier we use.

3.8.1 Computational Complexity

In the training phase, there is no difference between the paradigms. For each case, one simply has to estimate the parameters for the various density models involved. In the test phase, for all paradigms, the probabilities $p(\vec{x}|S_i, w_j)$ have to be computed. However the way in which these terms are manipulated is different. Moreover, the search for the optimal solution is different for the different cases. We will provide a brief theoretical analysis of the four paradigms. In particular, we will estimate the number of multiplies, adds and density computations[3] to calculate the score of each point in our search space. We will also indicate the total complexity of the search process.

- Paradigm 1:

  Number of Multiplies: $Ln + (N-1)nL$

  Number of Additions: $(N-1)nL$

  Number of Density Computations: NLn

  Search: There are L searches of $O(n)$ each.

- Paradigm 2: Here we perform a joint optimization and the number of multiplies, additions and the search complexity are all $O(n^L)$. However, in our implementation, we used a dynamic programming approach based on the $A*$ search [27]. The computational complexity was considerably reduced, but difficult to analyze in the same framework as the other paradigms. Consequently, we have not included an analysis of this paradigm in this thesis. Since the implementation of Paradigm 2 was done in C, an unbiased empirical comparison could not be done. We suspect, however, that this is the most expensive recognition paradigm.

---

[3]These are the computations involved in computing $p(\vec{x}|S_i, w_j)$.

- Paradigm 3:

  Number of Multiplies: $nLN + LN$

  Number of Additions: 0

  Number of Density Computations: NLn

  Search: There are NL searches of $O(n)$ each, and 1 search of $O(N)$.

- Paradigm 4:

  Number of Multiplies: $Ln + (N-1)nL + nLN + LN$

  Number of Additions: $(N-1)nL + (n-1)LN$

  Number of Density Computations: NLn

  Search: There are L searches of $O(n)$ each.

To empirically compare the different paradigms, we decided to measure the time taken to run the classification paradigms on the entire test set. For this purpose, we used the same training set of 16324 tokens, reducing dimensions using linear discriminant analysis, and the same test set as before. Clustering of speakers into 2, 4, 8, 16 and 32 clusters was done. In each case the total elapsed time for each paradigm to classify every token was measured and has been plotted in Figure 3.15. Paradigms 1, 3, and 4 were implemented entirely in S-Plus. In the case of Paradigm 2, however, the $A*$ search was written in C and was invoked from S-Plus.

From the figure, the time taken increases with increase in the value of N for each paradigm. This is not surprising, since the number of density computations and multiplies increases with N in each paradigm. Also, Paradigm 3, whose search depends more strongly on N than any other paradigm, seems to have a greater rise with N than the other paradigms. Notice here that Paradigm 2 which may actually be the most expensive computationally, is made much quicker using the $A*$ search and the C-routines which run faster. Paradigm 1 is the fastest which is also in accordance with our theoretical
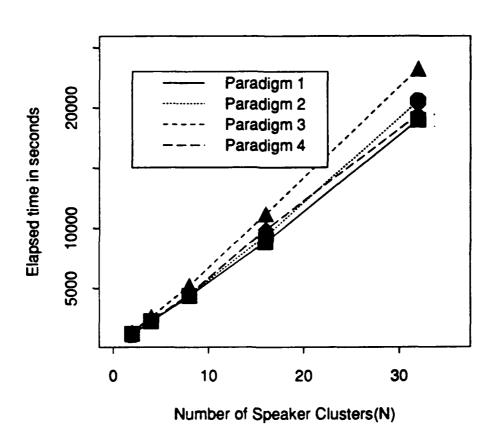
Figure 3.15: Elapsed time with number of speaker clusters for the different paradigms

predictions.

3.8.2 Classifier: Multilayer Perceptrons

Finally our last set of experiments in this chapter investigates the use of a different kind of classifier, viz., the multi-layer-perceptron (MLP). In all our previous examples, we had used a Bayesian classifier with Gaussian densities and diagonal covariance matrices. On a few occasions we have used a full covariance matrix but that has been the limit of our variation.

The particular MLP architecture for phonetic recognition has been described by Leung [18]. The MLP is found to have several characteristics which are particularly advantageous for phonetic classification tasks. Firstly, it does not make assumptions about the underlying probability distribution of the input data. Secondly, the MLP utilizes the training of connection weights to form decision regions, instead of using specific distance metrics (such as the Euclidean or Itakura [11]) to measure similarity. Very often the choice of distance metric is crucial for robustness in performance and may also put constraints on the input representation of a classifier. For example, discrimination by the Euclidean distance relies on differences in energy in the speech signal, and may be less suited for representations such as the synchronous response of SAM which has its energy information normalized. Thirdly, the MLP accepts both continuous inputs such as acoustic attributes and/or binary inputs like linguistic features. This allows us to integrate heterogeneous sources of information as an input representation. Finally, the MLP is capable of forming disjoint decision regions in the multi-dimensional input space for the same class without supervision. This may be especially suitable for modelling the various allophones of a phoneme or the different speaker realizations of the same phoneme.
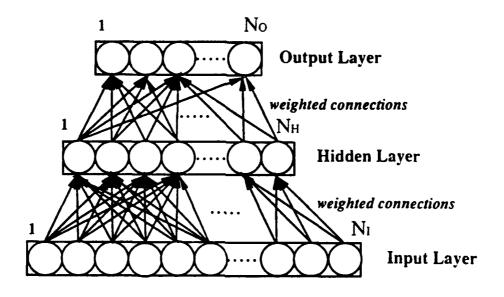
Figure 3.16: Structure of Multi-layer Perceptron

Network Structure

The network used has one hidden layer, and is illustrated in Figure 3.16. The number of output units N_0 depends on the number of classes to be recognized. The size of the network is determined by the number of units in the hidden layer, N_H. The number of input units depends upon the dimensionality of the input feature vector. It has been shown [8] [2] that training a neural network using a mean square error criterion gives network outputs that approximate posterior class probabilities. We want to see whether imposing speaker constraints will help in a neural network framework. For the

86

purpose of illustration, we picked Paradigm 1 and one speaker constraining paradigm viz. Paradigm 3 to compare against each other.

Shown in Figure 3.17, is the method by which we can coerce a multi-layer perceptron to yield the right probability measures to be used for Paradigms 1 and 3. Network A is trained only on male tokens and hence the output of the network gives $p(C|\vec{x}, S_1)$ where speaker type S_1 includes male speakers. The output of network B, which is trained on only female tokens, yields $p(C|\vec{x}, S_2)$ where S_2 includes female speakers. Networks A and B have 8 output units corresponding to the 8 vowel classes. Network C has 2 output units corresponding to the speaker groups (speaker gender in our case). This yields as its output $p(S_i|\vec{x})$ for any test token. Each network has as many input nodes as there are dimensions in the input feature vector. Further, each network has 32 hidden nodes.

Paradigm 1

Recall Paradigm 1 was

$$\max_{C_1,...,C_L} \prod_{j=1}^{L} p(C_j|\vec{x_j}) \tag{3.9}$$

This can be rewritten (with decomposition into speaker distributions) as

$$\max_{C_1,...,C_L} \prod_{j=1}^{L} \sum_{i=1}^{N} p(S_i, C_j|\vec{x_j}) \tag{3.10}$$

or, equivalently,

$$\max_{C_1,...,C_L} \prod_{j=1}^{L} \sum_{i=1}^{N} p(S_i|\vec{x_j})p(C_j|S_i, C_j) \tag{3.11}$$

All the terms in the above equation are *a-posteriori* probabilities and can be obtained from the three network structures.
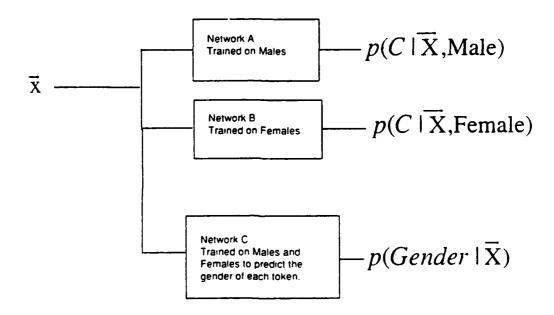
Figure 3.17: Arrangement of networks to implement Paradigms 1 and 3. The output of each network provides terms which can be suitably combined to obtain the optimizing expressions for the two paradigms.

Paradigm 3

Recall Paradigm 3 was

$$\max_{S_i, C_1, .., C_L} p(S_i, C_1, .., C_L | \vec{x}_1, .., \vec{x}_L) \qquad (3.12)$$

or equivalently

$$\max_{S_i, C_1, .., C_L} p(S_i | \vec{x}_1, .., \vec{x}_L) p(C_1, .., C_L | \vec{x}_1, .., \vec{x}_L, S_i) \qquad (3.13)$$

This can be rewritten as

$$\max_{S_i, C_1, .., C_L} p(S_i) p(C_1, .., C_L) p(\vec{x}_1, .., \vec{x}_L | C_1, .., C_L, S_i) \qquad (3.14)$$

Further,

$$\max_{S_i, C_1, .., C_L} p(S_i) \prod_{j=1}^{L} p(C_j) p(\vec{x}_j | C_j, S_i) \qquad (3.15)$$

which is the same as,

$$\max_{S_i, C_1, .., C_L} p(S_i) \prod_{j=1}^{L} p(\vec{x}_j | S_i) p(C_j | \vec{x}_j, S_i) \qquad (3.16)$$

Finally,

$$\max_{S_i, C_1, .., C_L} p(S_i) \prod_{j=1}^{L} \frac{p(S_i | \vec{x}_j)}{p(S_i)} p(C_j | \vec{x}_j, S_i) \qquad (3.17)$$

In the above equation we again have only *a-posteriori* terms which can be obtained from our networks. These are manipulated to yield **Paradigm 3** which imposes a speaker constraint.

Experiment 1

In this experiment, we used all our 16324 training tokens, each represented by a 40 dimensional vector which was the spectral average of the middle-

third of the vowel token. Thus the number of input nodes N_I was 40 and there were 8 output nodes corresponding to the 8 vowel classes. This was the structure of networks A and B. Network C had only 2 output nodes. Our results using Paradigm 1 was 57.38% and that using Paradigm 3 was 53.90%. This drop was disturbing. However, on closer investigation, we found that our gender network was not very sensitive. The dynamic range of the output *a-posteriori* probability was very poor and $p(S_i|\vec{x_j})$ always ranged from 0.48 to 0.52 for each token, although the gender of each token was identified correctly 91.44% of the time. Note that the *a-priori* probabilities for males and females were 0.67 and 0.33, respectively. Therefore, the term $\frac{p(S_i|\vec{x_j})}{p(S_i)}$ was very high for females, consistently biasing the maximization towards females and classifying every speaker on the basis of the female network. However, the female network having fewer training tokens was poorly trained and yielded only 53% accuracy. This was presumably the reason behind the poor performance of Paradigm 3.

Experiment 2

Nevertheless, we wanted to see if the whole idea of decomposing the overall population into speaker groups and imposing some kind of speaker constraints was meaningful in an MLP context, and so we abandoned our carefully constructed paradigms and resorted to a different method. We had two different networks trained on males and females and we had another network which merely predicted the gender of the speaker on the basis of his or her tokens. As we have seen above, although the shifts in the *a-posteriori* probabilities are very slight, they are nevertheless in the right direction and we can predict the gender of each token 91% of the time. We presented our gender network with all the tokens produced by the same speaker, and then on the basis of the gender predicted for each token, we classified the speaker as male

or female depending upon how many of the tokens were classified as each. This actually worked very well and of our 65 test speakers, we predicted the gender correctly in 64 cases. Depending upon the gender predicted, we accepted the output of either the male or the female network. This provided us with 60.71% accuracy in token classification. This was compared against a baseline where there was only one network trained on all the training tokens without distinguishing between male and female tokens. Our baseline was 59.5% and the difference was significant at the 0.01 level using McNemar's test.

The two experiments outlined above were not conclusive but they did provide an interesting dimension to the investigations of this thesis. Firstly, we found that a multilayer perceptron actually performed worse on our task than the Gaussian classifier. Admittedly, we did not experiment enough with the topology of the network or the number of training iterations to obtain peak performance. Furthermore, we also found that breaking our training speakers into groups and having separate networks trained on these, provided us with some gain in performance. This is in consonance with the ideas of this thesis. However the mathematical formulation can not be directly implemented using an MLP and considerable manipulation is needed to coerce the appropriate terms to perform the optimization. We found that our mathematical formulation was not very effective in this case due to the dynamic range problem described earlier.

3.8.3 Summary

This set of experiments investigated the issues of computational complexity and the kind of classifier used. The broad conclusions are reiterated:

- We provided a theoretical analysis of the four paradigms of recognition. We found that Paradigm 1 was computationally least expensive and Paradigm 2 was the most expensive. However the complexity of Paradigm 2 can be reduced by a dynamic programming approach, thus making a fair comparison impossible. We validated our theoretical predictions with an empirical comparison.

- We showed how a multi-layer perceptron could be used to coerce *a-posteriori* probabilities which could then be combined to perform recognition using Paradigms 1 and 3. However, we found a decrease in performance in going from Paradigm 1 to 3 and provided an explanation for the mechanism behind this. Altering our scheme slightly, we could, however, still impose speaker constraints using an MLP framework, resulting in improvement in performance.

3.9 Chapter Summary

This chapter investigated in detail, several factors related to the performance of our speaker-constraining paradigms (2,3 and 4) against a baseline (Paradigm 1) on a task of vowel classification. Various relevant issues are raised in Section 3.1 and investigated experimentally in Sections 3.6 through 3.8. In the next chapter we expand to a larger task and in the final chapter we reiterate the important results of this thesis.

Chapter 4

Phonetic Classification on a Task of All Phonemes

4.1 Motivation

So far we have developed several systematic techniques by which to enforce speaker constraints in phonetic classification. We have looked at several factors which might affect the relative success or failure of these schemes. The previous chapter concerned itself with a detailed experimental analysis of all these factors on a task of vowel classification. The next obvious extension is to look at a larger task. So we decided to compare our different paradigms of classification on a task of classifying all 39 phonemes of American English. Rather than repeat all the experiments of Chapter 3 on this larger task, we have decided to choose a particular set of model assumptions, token representations and speaker-group selection to validate our claim that imposing speaker constraints in phonetic classification leads to superior performance.

By the nature of their acoustic production, it seems intuitively clear that different sounds produced by the same speaker should be correlated. We already know that for vowels, speaker constraints result in superior classifi-

Vowels:	(i,ɪ,e,ɛ,æ,ɑ,ɔ,ʌ,ʊ,u, ɑʸ,ɔʸ,ɑʷ,ə,ɝ)
Semivowels:	(l,w,y,r)
Nasals:	(m,n,ŋ)
Fricatives:	(s,š,z,ž,f,θ,v, ð)
Stops:	(p,b,t,d,k,g)
Others:	(č,ǰ,h)

Table 4.1: Phonemes of American English.

cation accuracy. It is our intent to observe the extent to which this superior performance is affected by addition of other sound classes. Furthermore, we would like to observe the break-up of the total improvement in terms of improvement for different broad phonetic classes.

In the next section we will describe the experiment we performed which will be followed by a brief description of the results.

4.2 Experimental Set-Up

4.2.1 Task

The task was to classify the 39 phonemes of American English. These are grouped in terms of broad sound classes in Table 4.1

4.2.2 Corpus

The corpus used was TIMIT which has been used in all our experiments in this thesis. For reasons of consistency, the training set consists of the same 325 speakers used in the experiments of Chapter 3. Our test set consisted of the same 65 speakers as before. We used tokens excised from the SX and SI sentences of the corpus resulting in about 66,000 training tokens and 13,000

	Number of Speakers (M/F)	Number of Tokens
Training	325 (213/112)	66042
Test	65 (52/13)	13634

Table 4.2: Corpus used for experiments.

test tokens in all. A summary of the database is provided in Table 4.2.

4.2.3   Signal Processing

The speech signal is sampled at 16 kHz and a spectral vector is computed every 5 ms. Each frame produces a 40 dimensional spectral vector which is the output of an auditory model developed by Seneff [24]. This is exactly the same representation for speech as used earlier. For each each token, we obtained the spectral averages of the first, middle and last third of the token. These vectors were then concatenated to yield a 120 dimensional vector. This space was rotated using principal components analysis and reduced to 30 dimensions for the experiment which is reported here. We later adjusted the number of dimensions very slightly to improve our absolute performance.

4.2.4   Model Assumptions

Clearly for this experiment, n, the number of classes is equal to 39. We divided the population of training speakers into two supervised groups: Male and Female and so $N = 2$. Each speaker uttered approximately 200 tokens and so $L \sim 200$. Furthermore we assume our distributions are Gaussian with a diagonal covariance matrix.

Sound Class	Improvement (%)	No. of test tokens
Vowels	1.06	5021
Semivowels	1.10	1820
Nasals	0.98	1522
Fricatives	2.65	2606
Stops	0.74	2287
Others	0.00	378

Table 4.3: Improvements in percentage accuracy for different sound classes between Paradigm 1 and Paradigm 3.

4.3   Results and Discussion

For the task mentioned above we trained our models on the full training set using the conditions described in Section 4.2. We then performed classification of the test tokens using Paradigms 1 through 4. Paradigm 1 performed at 50.06% accuracy and Paradigms 2, 3 and 4 all yielded identical results at 51.33% classification accuracy. This difference was significant using Mc-Nemar's Test at the 0.001 level. As a matter of fact, p is of the order of 10^{-6}. This is encouraging because it means that our speaker constraining models continue to outperform the baseline even for this larger task. It is also worthwhile to observe that Paradigms 2, 3 and 4 yield identical results. We suspect that this is due to our choice of $N = 2$ and very large value of $L \sim 200$. Recall from Chapter 3 that as L increases, the speaker constraining paradigms start becoming more and more similar. Intuitively we see that Paradigms 2, 3 and 4 converge to the same performance in the asymptotic case of infinite tokens to optimize over for the joint assignment.

The overall improvement is 1.27%. The improvement for the different sound classes (obtained by decomposing the overall confusion matrix) is shown in the Table 4.3.

This experiment confirms the fact that speaker constraints help for all phonetic classes. However, due to simplistic representations and poor model assumptions, our absolute performance is rather low. We changed our model assumptions to full covariance Gaussian distributions and used 35 principal components instead of 30 and this resulted in a baseline performance (Paradigm 1) of 55.05% and 56.21% for the other paradigms. It has been observed that hair-cell representations are not very Gaussian. Using a different representation would presumably bolster performance even more, as others have found empirically [9].

Chapter 5

Conclusions

In speech recognition one tries to find an optimal mapping from the acoustic to the lexical domain. In this thesis we have tried to explicitly model two features into this mapping process. Firstly we have tried to incorporate speaker-specific models to try and capture the inter-speaker variability. Secondly and more importantly, we have argued that different sounds produced by the same speaker are correlated and hence the acoustic-to-lexical mapping should be done jointly (rather than individually) for all sounds produced by the same test speaker. This is equivalent to applying speaker constraints in classification.

5.1   Results of This Thesis

We have developed several systematic techniques of classification which impose speaker constraints. The baseline (Paradigm 1) incorporates speaker variability but applies no speaker constraints at all. Paradigms 2, 3, and 4 impose speaker constraints in slightly different ways. We compared these paradigms on a task of vowel classification and our broad conclusions are reiterated here:

5.1.1 Supervised Clustering of Speakers Into Groups on the Basis of Gender

When $N = 2$ and the two speaker groups are males and females, we are in effect imposing gender-constraints. We observe:

- Paradigms 2, 3, and 4 (i.e. speaker constraining paradigms) outperform Paradigm 1 given sufficient training data. This difference in performance is significant.

- Paradigms 2,3 and 4 do not differ significantly in performance from one other.

- The above result holds true for various representations for the vowel tokens.

- As L, the number of tokens used for optimization, increases, the difference between the speaker constraining paradigms and the baseline increases too. For $L = 1$ (equivalent to independence assumption between test tokens) the difference is insignificant.

- We experimented with a classifier based on multi-layer-perceptrons and found that with a little bit of modification, the speaker constraining paradigms again yielded significantly higher classification accuracy.

- We expanded our task to classification of 39 phonemes of American English and found significant improvement over the baseline on applying speaker constraints.

5.1.2 Unsupervised Clustering of Speakers Into Groups

We looked at several different ways to cluster our speakers into speaker groups. We found

- The success of speaker constraining paradigms depended upon how we clustered our speakers into speaker groups.

- The performance of speaker constraining paradigms showed a distinct peak with the number of clusters, N. The value of N at which the peak occurs was observed to be a function of the amount of training data used.

- We compared the computational complexity (measured as total run time for classification) for each of the paradigms and found Paradigm 1 to be the fastest and Paradigm 3 to be the most expensive. Paradigm 2 was implemented in C and so a fair comparison could not be made with the other paradigms.

The above seems to suggest that different sounds produced by the same speaker are indeed correlated and exploiting these correlations in phonetic classification leads to potential improvement in classification accuracy.

5.2 Limitations and Future Work

Finally, we will conclude with some of the limitations of this work. We will also provide suggestions for further improvement to overcome those limitations and expand the scope of this thesis. Wherever appropriate, we have also included comparisons to other work done in similar areas.

5.2.1 Absolute Performance

We have obtained improvements by applying speaker constraints and these are comparable to other similar schemes. For example, [20] implemented parallel male and female recognizers exactly like Paradigm 3 in our case and obtained similar improvements. However, our absolute performance is much

100

poorer. Further, [26] have done updating of models using MAP estimates much in the same fashion as our Paradigm 4. Again, although improvements are comparable, our absolute performance is worse. Finally, our overall classification accuracy for vowels and for all phonemes is lower than the best results obtained by [19] and [18]. This could be due to several reasons:

- Representation: All our representations were based on the spectral vectors computed from Seneff's Auditory Model. These have been found to be sometimes markedly non-Gaussian in distribution [9]. Since we largely used Gaussian models, this might have reduced performance. Furthermore, in some cases, as in our vowel experiments, we made measurements only on the middle-third of our tokens which might well have been insufficient. On the issue of representation, it is also noteworthy that we want to extract features which maximally characterize speaker and phonetic identity. Further work can be done on the kinds of features which do this best. Features for extracting phonetic identity alone have been investigated by [19].

- Context: Our task consisted of phonemes in varying contexts. However, we had no context modelling at all in our system. This would surely have reduced recognition performance. For example, [26] dealt with isolated alphabets where context had less influence, and their absolute performance was superior to ours. Some more work could be done to incorporate context models in our theoretical framework. This can easily be done but might involve considerable computational expense in implementation.

- Classifier: We used simple Gaussian classifiers. This, coupled with the non-Gaussian nature of the measurements we made might have hurt us.

5.2.2 Expansion to Isolated Word and Continuous Speech Recognition

Our mathematical formulation was general so that the pattern classes w_i need not necessarily refer to phonemes. It was only in the empirical comparisons that we used vowels first and later phonemes of American English. Considerable work could be done in expanding the ideas of this thesis to isolated work or continuous speech recognition. There are several ways in which this could be done. For example, for isolated word recognition, we might redefine the w_i's to refer to individual words. In that case, since different words have different temporal structures, we might have problems in time-normalizing them and obtaining a vector $\bar{x}$ of the same dimension for each word. Simple averaging, as in the case of phonemes, might prove to be insufficient. Alternatively, we might decide to drive a word-recognition system with a phonetic recognizer and a suitable framework for this will have to be devised. Similar issues will be involved in continuous speech recognition. Finally, some theoretical work could be done to relax the probabilistic interpretation of our paradigms of recognition and to extend the same idea to other score-based schemes of recognition. This will add a lot of flexibility to the theoretical framework. We have experimented with this idea when trying to change our classifier to a multi-layer perceptron but much more work could be done.

5.3 Summary

In this chapter, we have reiterated the core idea of this thesis viz. that different sounds produced by the same speaker are correlated and exploiting this correlation could lead to potential improvement in speech recognition. We have developed systematic ways of doing this and our findings are summarized here. We have also discussed some shortcomings and suggested further

areas of investigation.

Bibliography

[1] Becker, R.A., Chambers, J.M., and Wilks, A.R., *The New S Language,* a Wadsworth and Brooks/Cole publication, 1988.

[2] Boulard, H., and Wellekens, C., "Speech Pattern Discrimination and Multi-layer Perceptrons," *Manuscript M.211,* Phillips Research Lab., Brussels, Belgium.

[3] Choukri, K., and Chollet, G., "Adaptation of automatic speech recognizers to new speakers using canonical correlation analysis techniques," *Computer Speech and Language,* Vol. 1, pp 95-107, 1986.

[4] Duda, R., and Hart, P., *Pattern Classification and Scene Analysis,* a Wiley-Interscience publication, 1973.

[5] Feng, M., Kubala, F., Schwartz, R., and Makhoul, J., "Improved Speaker Adaptation Using Text Dependent Spectral Mappings," *Proc. of the International Conference on Acoustics, Speech and Signal Processing,* pp. 131-134, 1988.

[6] Gallagher, R.G., *Information Theory and Reliable Communication,* John Wiley and sons, 1968.

[7] Gillick, L., and Cox, S.J., " Some Statistical Issues in the Comparison of Speech Recognition Algorithms," *Proc. of the International Conference on Acoustics, Speech and Signal Processing,* pp. 532-535, April, 1986.

[8] Gish, H., "A Probabilistic Approach to the Understanding and Training of Neural Network Classifiers," *Proc. of the International Conference on Acoustics,Speech and Signal Processing*, pp. 1361-1364, April 1990.

[9] Glass, J.R., Personal Communication

[10] Hogg, R.V., and Tanis, E.A., *Probability and Statistical Inference,* 3rd edition, Macmillan Publishing Company, 1988.

[11] Itakura, F., "Minimum Prediction Residual Principle Applied to Speech Recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing,* February 1975.

[12] Johnson, R.A., and Wichern, D.W., *Applied Multivariate Statistical Analysis,* Second Edition, a Prentice Hall publication, 1988.

[13] Kucera, H., and Francis, W.N., *Computational Analysis of Present-Day American English,* Brown University Press, Providence, R.I., 1967.

[14] Lamel, L., Kassel, R., and Seneff, S., "Speech Database Development: Design and Analysis of the Acoustic-Phonetic Corpus," *Proc. DARPA Speech Recognition Workshop,* Report No. SAIC-86/1546, pp. 100–109, February 1986.

[15] Lee, C.H., Rabiner, L.R., Pieraccini, R., and Wilpon, J.G., "Acoustic modeling for large vocabulary speech recognition," *Computer Speech and Language Processing,* Vol. 4, No. 2, April 1990.

[16] Lee, K.F., *Large Vocabulary Speaker Independent Continuous Speech Recognition: The Sphinx System,* Doctoral Thesis, Carnegie Mellon University, Pittsburgh, PA, 1988.

[17] Lee, K.F., Hon, H.W., and Reddy, D.R., "An Overview of the SPHINX Speech Recognition System," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, pp. 35-45, January 1990.

[18] Leung, H.C., *The Use of Artificial Neural Networks in Phonetic Recognition*, Doctoral Thesis, Massachussetts Institute of Technology, May 1989.

[19] Meng, H.M., *The Use of Distinctive Features for Automatic Speech Recognition*, S.M. Thesis, Massachussetts Institute of Technology, May 1991.

[20] Murveit, H., Weintraub, M., and Cohen, M., "Training Set Issues in SRI's DECIPHER Speech Recognition System," *Proc. Third DARPA Speech and Natural Language Workshop*, pp. 337-340, Hidden Valley, PA, June, 1990.

[21] Myers, R.H., *Classical and Modern Regression with Applications*, PWS Publishers, 1986.

[22] Rabiner, L.R., "On Creating Reference Templates for Speaker Independent Recognition of Isolated Words," *IEEE Trans. ASSP*, Vol. ASSP-26, No. 1, pp. 34–42, 1978.

[23] Schwartz, R., Chow, Y., Kimball, O., Roucos, S., Krasner, M., and Makhoul, J., "Context-Dependent Modelling for Acoustic-Phonetic Recognition of Continuous Speech," *Proc. of the International Conference on Acoustics, Speech and Signal Processing*, pp. 1205-1208, August 1985.

[24] Seneff, S., "A Joint Synchrony/Mean Rate Model of Auditory Speech Processing." *Journal of Phonetics*, Vol.16, No.1, pp. 55-76, 1988.

[25] Shikano, K., Lee, K.F., and Reddy, D.R., "Speaker Adaptation Through Vector Quantization," *Proc. of the International Conference on Acoustics, Speech and Signal Processing,* pp. 2343-2346, 1986.

[26] Stern, R.M., and Lasry, M.J., "Dynamic Speaker Adaptation for Feature-Based Isolated Word Recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing,* June, 1987.

[27] Winston, P.H., *Artificial Intelligence,* Addison Wesley, 2nd Edition, 1984.

[28] Zue, V.W., Glass, J.R., Phillips, M., and Seneff, S., "Acoustic Segmentation and Phonetic Classification in the Summit System," *Proc. of the International Conference on Acoustics,Speech and Signal Processing,* pp. 389-392, May 1989.